Traitement automatique des langues

Réseaux sociaux

sous la direction de
Atefeh Farzindar
Mathieu Roche

Vol. 54- n°3/ 2013

ATALA

Réseaux sociaux

ATALA

Revue de
l'Association
pour le Traitement
Automatique
des Langues

TAL
Vol.
54

n°3
2013

Réseaux sociaux

ATALA

Traitement automatique des langues

Revue publiée depuis 1960 par l'Association pour le Traitement Automatique des LAngues (ATALA), avec le concours du CNRS, de l'Université Paris VII et de l'Université de Provence

ISSN 1965-0906

http ://www.atala.org/-Revue-TAL-

Traitement automatique des langues

Comité de rédaction

Rédacteurs en chef

Membres

Traitement automatique des langues

Volume 54 – n°3/2013

TRAITEMENT AUTOMATIQUE DU LANGAGE NATUREL POUR L'ANALYSE DES RÉSEAUX SOCIAUX (TAL ET RÉSEAUX SOCIAUX)

Table des matières

. . . . . .

Les défis de l'analyse des réseaux sociaux pour le traitement automatique des langues

Atefeh Farzindar[*] [**], **Mathieu Roche**[***] [****]

[*] *NLP Technologies Inc., Montréal (Québec), Canada*
www.nlptechnologies.ca
farzindar@nlptechnologies.ca
[**] *Université de Montréal, Montréal (Québec), Canada*
[***] *UMR TETIS (Cirad, Irstea, AgroParisTech), Montpellier, France*
mathieu.roche@cirad.fr
[****] *LIRMM (CNRS, Université de Montpellier), Montpellier, France*
mathieu.roche@lirmm.fr

RÉSUMÉ. *Les réseaux sociaux intègrent un volume et une variété sans précédent de données textuelles. Leur analyse permet de mieux comprendre des comportements sociaux et certaines évolutions sociétales. L'étude des messages échangés, qui sont par nature complexes, représente de nouvelles problématiques pour le traitement automatique des langues (TAL). Dans ce contexte, cet article introductif au numéro spécial de la revue TAL présente les défis liés à l'infobésité des données issues des réseaux sociaux puis discute de l'utilisation des méthodes de TAL pour traiter le contenu textuel de ces nouveaux modes de communication.*

ABSTRACT. *Social networks incorporate an unprecedented amount and variety of textual data. The analysis of this information furthers our understanding of social behaviors and some trends. The study of inherently complex messages sent between users represents new problems for Natural Language Processing (NLP). In this context, the first article in this special issue of the TAL journal introduces the challenges of information overload from social networks, and discusses the use of NLP methods for processing the textual content of these new modes of communication.*

MOTS-CLÉS : *TAL, réseaux sociaux, analyse sémantique.*

KEYWORDS: *NLP, social networks, semantic analysis.*

TAL. Volume 54 – n° 3/2013, pages 7 à 16

1. Introduction

Les réseaux sociaux, structures dynamiques formées d'individus ou d'organisations, ont toujours joué un rôle majeur dans nos sociétés. Ils se sont développés et diversifiés avec le Web 2.0 qui ouvre la possibilité aux utilisateurs de créer et de partager du contenu par l'intermédiaire de multiples plates-formes (blogues, microblogues, wikis, sites de partage, etc.). Ces modes de communication sont de puissants outils collectifs où s'invente et s'expérimente le langage. De nouveaux sens sont alors associés à certains mots ou syntagmes et la création de mots ou de nouvelles structures syntaxiques se généralise. La création, la dissémination et le traitement du matériau textuel issu des réseaux sociaux sont discutés dans ce numéro spécial. Plus globalement, cet article et ce numéro spécial permettent de mettre en exergue une nouvelle manière de communiquer illustrée par (1) l'article « Code-Mixing in Social Media Text : The Last Language Identification Frontier ? » de Amitava Das et Björn Gambäck sélectionné parmi sept articles soumis et (2) l'article invité « Détection d'évènements à partir de Twitter » de Houssem Eddine Dridi et Guy Lapalme.

Pour traiter les masses de données issues des réseaux sociaux aujourd'hui disponibles (c'est-à-dire l'infobésité), la problématique de recherche du « Big Data » est classiquement mise en avant avec les trois V qui la caractérisent : volume, variété et vélocité. Cet article discute, dans un premier temps, de ces trois caractéristiques appliquées aux réseaux sociaux (section 2). Puis, dans le cadre de l'infobésité décrite de manière générale, nous étudierons, en section 3, la manière d'analyser le contenu des messages issus des réseaux sociaux par des méthodes de traitement automatique des langues (TAL). En effet, certaines métadonnées (par exemple, les hashtags) et les descripteurs linguistiques (ou unités lexicales) issus des messages constituent un socle solide pour l'analyse des réseaux sociaux. Ils permettent de mettre en avant différentes communautés socio-économiques, politiques, géographiques, etc. Par ailleurs, les descripteurs linguistiques sous forme de mots ou syntagmes permettent d'analyser avec précision les sentiments et opinions contenus dans les messages. Par exemple, les spécificités lexicales, graphiques voire syntaxiques (émoticônes, abréviations, répétition de caractères, etc.) véhiculent des informations précieuses pour l'analyse de sentiment (détection fine des émotions, identification de l'ironie, etc.).

2. Infobésité et analyse des réseaux sociaux

Au cœur de la structure des réseaux sociaux se trouvent des acteurs (personnes ou organisations) reliés entre eux par un ensemble de relations binaires (par exemple, liens ou interactions). Dans ce contexte, le but est de modéliser la structure d'un groupe social, en vue de déterminer l'influence qu'elle exerce sur d'autres variables, et d'assurer le suivi de son évolution. L'analyse sémantique des médias sociaux (ASMS) se définit comme l'art de comprendre comment on recourt aux réseaux sociaux pour générer du renseignement sociétal, stratégique, opérationnel ou tactique. Récemment, des ateliers tels que « L'analyse sémantique des médias sociaux » et « L'analyse linguistique dans les médias sociaux » de EACL 2012 (Farzindar et Inkpen, 2012),

NACL-HLT 2013 (Farzindar *et al.*, 2013) et EACL 2014 (Farzindar *et al.*, 2014) témoignent de l'intérêt grandissant à l'égard de l'impact des médias sociaux sur la vie quotidienne des individus, tant sur le plan personnel que professionnel. L'ASMS favorise la création d'outils et d'algorithmes visant à surveiller, à saisir et à analyser les données des médias sociaux qui sont volumineuses (section 2.1), produites en temps réel (section 2.2) et de nature hétérogène (section 2.3).

2.1. *Volume*

Un rapport publié par eMarketer (New Media Trend Watch, 2013) estimait qu'une personne sur quatre à l'échelle mondiale était susceptible d'utiliser les médias sociaux en 2013. Les statistiques sur les médias sociaux pour l'année 2012 révèlent que Facebook a dépassé la barre des huit cents millions d'utilisateurs actifs, dont deux cents millions de nouveaux adhérents au cours d'une seule année. La plate-forme Twitter, quant à elle, compte maintenant cent millions d'utilisateurs et LinkedIn, plus de soixante-quatre millions, en Amérique du Nord seulement (Digital Buzz, 2012). À titre d'exemple, plus de trois cent millions de tweets seraient envoyés à Twitter chaque jour (Tang *et al.*, 2014).

L'analyse et la veille de ce riche contenu sans cesse renouvelé donnent accès à une information précieuse que les médias traditionnels ne peuvent fournir (Melville *et al.*, 2009). L'analyse sémantique des médias sociaux a ouvert la voie à l'analyse de données volumineuses, discipline émergente inspirée de l'apprentissage automatique, de l'exploration de données, de la recherche documentaire, de la traduction automatique, du résumé automatique et du TAL plus globalement.

2.2. *Vélocité*

Les données issues des réseaux sociaux sont en général produites en temps réel. Par ailleurs les messages traitant d'un sujet commun peuvent véhiculer des émotions, des néologismes ou des rumeurs. Ces messages peuvent provenir de localisations différentes qu'il est nécessaire de prendre en compte dans le cadre de la vélocité des données.

Les médias sociaux soulèvent l'important problème de la recherche d'événements en temps réel et de la nécessité de les détecter (Farzindar et Wael, 2015). L'objectif de la recherche documentaire dynamique et de la recherche d'événements en temps réel est de mettre en place des stratégies de recherche efficaces à partir de différentes fonctionnalités qui tiennent compte de multiples dimensions, y compris les liens spatiaux et temporels (Gaio *et al.*, 2012 ; Moncla *et al.*, 2014). En outre, les discussions propres à un événements peuvent mêler, sur une période très courte, différents sujets parfois écrits en différentes langues. Ce point illustre la problématique liée à l'hétérogénéité des données qui est détaillée dans la section suivante.

2.3. *Variété*

L'importante quantité d'informations accessible dans les médias sociaux représente une manne de renseignements. Mais les textes, rédigés par des auteurs différents dans une variété de langues et de styles, n'adoptent, en général, aucune structure précise et se présentent sous une multitude de formats : blogues, microblogues, forums de discussion, clavardages, jeux en ligne, annotations, classements, commentaires et FAQ générées par des utilisateurs, etc. Les variations sur le plan du contenu et du style rendent l'analyse globale difficile. De manière concrète, les applications décrites ci-dessous montrent la variété des domaines et des tâches menées à partir des réseaux sociaux.

2.3.1. *Secteur industriel*

L'intérêt pour la surveillance de données extraites des médias sociaux est considérable dans le secteur industriel. En effet, ces données sont susceptibles d'aider en optimisant de manière importante l'efficacité de la veille stratégique. L'intégration de telles données aux systèmes de veille stratégique déjà en place permet aux entreprises d'atteindre différents objectifs, notamment concernant la stratégie de marque et la notoriété, la gestion des clients actuels et potentiels et l'amélioration du service à la clientèle. Le marketing en ligne, la recommandation de produits et la gestion de la réputation ne sont que quelques exemples d'applications concrètes de l'ASMS.

2.3.2. *Défense et sécurité nationale*

Ce secteur s'intéresse en particulier à l'étude de ce type de sources d'information pour comprendre différentes situations, procéder à l'analyse des sentiments d'un groupe de personnes partageant des intérêts communs et rester vigilant aux menaces potentielles dans les domaines cibles. Certaines méthodes d'extraction d'information (par exemple l'extraction des entités nommées et des liens entre ces dernières) à partir du Web 2.0 sont souvent développées pour analyser le contenu des réseaux sociaux au sein desquels évoluent des utilisateurs mais aussi des organisations. De telles informations offrent de précieux renseignements en matière de sécurité nationale.

2.3.3. *Soins de santé*

Les forums de discussion qui sont des espaces d'échanges asynchrones de messages textuels sont très prisés par certains patients. En effet, ils sont associés à un véritable espace de liberté du discours. Ainsi, l'utilisation de Twitter ou des forums comme des plates-formes de discussion sur des sujets tels que les maladies, les traitements, les médicaments ou les recommandations à l'intention des professionnels et des bénéficiaires (patients, familles et aidants) illustre bien la pertinence des médias sociaux dans ce domaine. Par ailleurs, dans ce contexte éminemment subjectif, la caractérisation et la compréhension des perceptions que les patients ont de leur maladie et du suivi médical représentent un enjeu sociétal particulièrement intéressant pour les professionnels de santé (Bringay *et al.*, 2014 ; Abdaoui *et al.*, 2014).

2.3.4. *Politique*

La veille des médias sociaux permet d'assurer le suivi des mentions faites par différents citoyens d'un pays ainsi que de l'opinion à l'égard d'un parti politique. Le nombre d'abonnés que compte un parti est essentiel au déroulement de sa campagne électorale. L'extraction d'opinions et le suivi des déclarations publiées sur les réseaux sociaux permettent à un parti politique de mieux saisir la teneur de certains événements, lui donnant ainsi l'occasion de s'ajuster pour améliorer ses positionnements politiques voire ses propositions (Bouillot *et al.*, 2012 ; Bakliwal *et al.*, 2013).

Face au volume, à la vélocité et à la variété des données textuelles issues des réseaux sociaux, les méthodes de TAL à appliquer et à proposer se révèlent cruciales. Les nouveaux défis adressés au TAL dans un tel contexte sont détaillés dans la section suivante.

3. Les défis liés au traitement du contenu des réseaux sociaux

L'information diffusée dans les médias sociaux, notamment dans les forums de discussion, les blogues et les gazouillis, est riche et dynamique. L'application des méthodes habituelles de TAL dans ce contexte ne se fait pas sans difficulté en raison du bruit et de l'orthographe « inhabituelle ». L'importance des médias sociaux émane du fait que chaque utilisateur est désormais un auteur potentiel et que le langage se rapproche davantage de sa réalité que d'une quelconque norme linguistique (Zhou et Hovy, 2006). Les blogues, les gazouillis et les mises à jour de statuts sont rédigés de manière informelle, sur le ton de la conversation, et ressemblent plus à un « état d'âme » qu'au travail réfléchi et révisé avec le soin habituellement attendu d'un média papier. Ce caractère informel engendre différents défis au domaine du TAL.

Les outils du TAL conçus pour les données traditionnelles se heurtent, par exemple, à l'emploi irrégulier, voire l'omission, de la ponctuation et des majuscules. Une telle situation complique la détection des limites d'une phrase qui constitue une tâche de base essentielle pour l'analyse des textes. Par ailleurs, l'utilisation de binettes, l'orthographe incorrecte ou inhabituelle et la multiplication d'abréviations populaires compliquent les tâches telles que la segmentation et l'étiquetage morphosyntaxique. Une adaptation des outils traditionnels est nécessaire pour prendre en compte les nouvelles variations comme la répétition des lettres (par exemple, *suuuuuper*) (Hangya *et al.*, 2013). Un autre obstacle à toute forme d'analyse syntaxique est la grammaticalité, ou plutôt son absence fréquente dans les médias sociaux (Kong *et al.*, 2014). En effet, les phrases fragmentées sont devenues la norme à l'instar des phrases complètes et le choix entre différents homophones semble arbitraire (par exemple, *c'est*, *ces*, *ses*).

Outre ces aspects liés aux spécificités lexicales voire syntaxiques du contenu des messages échangés sur les réseaux sociaux, ces derniers génèrent beaucoup plus de bruit que les médias dits traditionnels. En effet, les réseaux sociaux comportent un

nombre considérable de pourriels, de publicités et une importante quantité de contenus non sollicités, non pertinents ou dérangeants. En outre, une grande partie du contenu qualifié d'authentique et de légitime ne répond pas mieux aux besoins d'information, et est donc jugée non pertinente, comme l'illustrait bien l'étude rapportée dans (André *et al.*, 2012), visant à mesurer la valeur que les utilisateurs accordent à différents gazouillis. Des quarante mille évaluations de gazouillis recueillies, 36 % recevaient la mention « vaut la peine d'être lu » et 25 %, « ne vaut pas la peine d'être lu ». Les gazouillis qui attestent seulement de la présence d'un utilisateur sur la plate-forme (par exemple, *Alloooo Twitter !*) se sont vus attribuer la plus faible valeur. Cela souligne l'importance du prétraitement, visant à filtrer les pourriels et autres contenus non pertinents, et de la création de modèles de gestion du bruit efficaces, en vue du traitement du langage dans les médias sociaux.

De nombreux domaines d'application qui prennent en compte ces caractéristiques propres aux réseaux sociaux sont alors étudiés comme par exemple, le résumé automatique, la détection d'événements et l'analyse de sentiments. Ces trois domaines de recherche appliqués aux réseaux sociaux et caractérisés par les trois V du « Big Data » sont détaillés ci-dessous.

– Comme illustré en section 2.1, avec la présence de textes courts, bruités et en nombre important, les médias sociaux se prêtent difficilement aux approches de TAL comme le résumé automatique. À titre d'exemple, les gazouillis, avec leur limite de cent quarante caractères, sont plus pauvres sur le plan contextuel que les documents traditionnels. Aussi, la redondance est problématique dans une suite de gazouillis, en partie en raison de la fonction de partage. Les expériences présentées dans (Sharifi *et al.*, 2010) avec les techniques d'exploration de données visant à générer des résumés automatiques de sujets à la mode sur Twitter les ont amenés à identifier l'important problème posé par la redondance de l'information. En outre, l'information diffusée dans les médias sociaux est hautement dynamique et caractérisée par l'interaction entre différents participants. Si elle complexifie d'autant plus le recours aux approches traditionnelles de résumés automatiques, elle offre en revanche l'occasion d'utiliser de nouveaux contextes pour enrichir les résumés, et permet même de créer de nouveaux procédés de résumés automatiques. Par exemple, l'article de (Hu *et al.*, 2007) suggère de procéder au résumé automatique d'une publication tirée d'un blogue en extrayant des phrases représentatives à partir d'informations recueillies de commentaires d'utilisateurs. (Chua et Asur, 2012) se concentrent, quant à eux, sur la corrélation temporelle de gazouillis pour extraire ceux susceptibles d'être pertinents pour le résumé automatique. Enfin, outre le contenu des messages, d'autres approches exploitent les informations liées à l'interaction entre utilisateurs pour produire un résumé des différents échanges (Lin *et al.*, 2009).

– Comme évoqué en section 2.2, l'identification d'événements dans les flux de données est une tâche particulièrement difficile (Allan, 2002). Dans le cadre de l'étude des réseaux sociaux, l'un des défis majeurs est la distinction entre l'information triviale et « polluée » et les événements concrets d'intérêt. La dispersion des

données, l'absence de contexte et la diversité du vocabulaire rendent les techniques traditionnelles d'analyse textuelle difficilement applicables aux gazouillis (Metzler *et al.*, 2007). En outre, différents événements n'atteindront pas la même popularité chez les utilisateurs et peuvent grandement varier sur le plan du contenu, de la période couverte, de la structure inhérente, des relations causales, du nombre de messages générés et du nombre de participants (Nallapati *et al.*, 2004).

– Alors que les médias traditionnels visent, en général, à diffuser une information objective, neutre et factuelle, les médias sociaux sont beaucoup plus porteurs de sentiments voire d'émotion (Neviarouskaya *et al.*, 2011 ; Bringay *et al.*, 2014) (cf. section 2.3). L'information subjective joue donc un rôle essentiel dans l'analyse sémantique des textes issus des réseaux sociaux. L'identification de sentiments repose généralement sur deux familles d'approches. La première est fondée sur des méthodes classiques d'apprentissage supervisé qui proposent des résultats tout à fait satisfaisants pour l'analyse de sentiments (Pang *et al.*, 2002). La seconde s'appuie sur des informations statistiques liées au nombre de descripteurs linguistiques positifs et négatifs qui apparaissent dans chaque texte (Turney, 2002). Dans le cadre de ces approches, il est alors pertinent d'utiliser des ressources existantes telles que SentiWordNet (Esuli et Sebastiani, 2006). Chaque caractéristique de cette ressource est associée à des scores numériques décrivant l'intensité des descripteurs linguistiques selon trois critères : objectif, positif et négatif. Notons que certaines approches récentes se concentrent sur l'identification des émotions associées aux descripteurs spécifiques des réseaux sociaux (par exemple, les hashtags issus des tweets) (Qadir et Riloff, 2014).

4. Conclusion

Les médias sociaux se définissent par le recours à des outils électroniques et à l'Internet dans le but de partager et d'échanger efficacement de l'information et des expériences (Moturu, 2009). Ils donnent accès à une information riche et sans cesse renouvelée que les médias traditionnels ne fournissent pas (Melville *et al.*, 2009). Les deux réseaux sociaux les plus populaires, Facebook et Twitter, sont étudiés dans les articles retenus de ce numéro spécial. Le premier (article de Amitava Das et Björn Gambäck) s'intéresse à l'identification des langues (anglais, bengali et hindi) que nous pouvons retrouver dans une même phrase ou un même message. En effet, comme évoqué dans cet article introductif, l'aspect multilingue associé aux réseaux sociaux demeure une problématique éminemment complexe. Outre le mélange des langues, les messages des réseaux sociaux ont des caractéristiques comme la présence de hashtags qui doivent aussi être étudiés dans les différentes applications. Ces aspects sont pris en compte dans un processus global de détection d'événements proposé dans l'article invité de ce numéro spécial (article de Houssem Eddine Dridi et Guy Lapalme).

Ce numéro spécial montre de quelle manière les méthodes de TAL contribuent à l'analyse des réseaux sociaux. Différents systèmes qui gèrent le contenu des forums

de discussion, des blogues et des microblogues, ont récemment connu des améliorations qui favorisent tant la formation de communautés virtuelles que la connectivité et la collaboration entre les utilisateurs (Osborne *et al.*, 2014). Alors que les médias traditionnels – tels que journaux, télévisions et radios – se caractérisent par un mode de communication unidirectionnel de l'entreprise jusqu'au consommateur, les médias sociaux, eux, proposent différentes plates-formes où l'interaction dans les deux sens est possible. Pour cette raison, ils représentent une source primaire d'information au moment de réaliser une veille stratégique. C'est ainsi que plus récemment, les recherches se sont concentrées sur l'analyse du langage dans les médias sociaux pour comprendre les comportements sociaux et concevoir des systèmes socioadaptés. L'objectif est d'analyser le langage dans une démarche pluridisciplinaire mêlant par exemple, informatique, linguistique, sociolinguistique et psycholinguistique (Brézillon *et al.*, 2013 ; Aiello et McFarland, 2014).

Remerciements

Nous remercions les auteurs pour la qualité des contributions, les relecteurs pour l'évaluation des articles soumis et Jean-Luc Minel pour son soutien et ses conseils avisés tout au long du processus.

5. Bibliographie

Abdaoui A., Azé J., Bringay S., Grabar N., Poncelet P., « Analysis of Forum Posts Written by Patients and Health Professionals », *Proceedings of European Medical Informatics Conference (MIE)*, p. 1185, 2014.

Aiello L. M., McFarland D. A. (eds), *Social Informatics - 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings*, vol. 8851 of *Lecture Notes in Computer Science*, Springer, 2014.

Allan J. (ed.), *Topic Detection and Tracking : Event-based Information Organization*, Kluwer Academic Publishers, Norwell, MA, USA, 2002.

André P., Bernstein M., Luther K., « Who Gives a Tweet ? : Evaluating Microblog Content Value », *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, p. 471-474, 2012.

Bakliwal A., Foster J., van der Puil J., O'Brien R., Tounsi L., Hughes M., « Sentiment Analysis of Political Tweets : Towards an Accurate Classifier », *Proceedings of the Workshop on Language Analysis in Social Media*, Association for Computational Linguistics, Atlanta, Georgia, p. 49-58, June, 2013.

Bouillot F., Hai P. N., Béchet N., Bringay S., Ienco D., Matwin S., Poncelet P., Roche M., Teisseire M., « How to Extract Relevant Knowledge from Tweets ? », *Information Search, Integration and Personalization - International Workshop, ISIP 2012, Springer, Revised Selected Papers*, p. 111-120, 2012.

Brézillon P., Blackburn P., Dapoigny R. (eds), *Modeling and Using Context - 8th International and Interdisciplinary Conference, CONTEXT 2013, Annecy, France, October 28 -31, 2013, Proceedings*, vol. 8175 of *Lecture Notes in Computer Science*, Springer, 2013.

Bringay S., Kergosien E., Pompidor P., Poncelet P., « Identifying the Targets of the Emotions Expressed in Health Forums », *Proceedings of Computational Linguistics and Intelligent Text Processing - 15th International Conference, CICLing 2014, LNCS, Part II*, p. 85-97, 2014.

Chua F. C. T., Asur S., Automatic Summarization of Events from Social Media, Technical report, HP Labs, 2012.

Esuli A., Sebastiani F., « SENTIWORDNET : A Publicly Available Lexical Resource for Opinion Mining », *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC)*, p. 417-422, 2006.

Farzindar A., Gamon M., Inkpen D., Nagarajan M., Danescu-Niculescu-Mizil C. (eds), *Proceedings of the Workshop on Language Analysis in Social Media*, Association for Computational Linguistics, Atlanta, Georgia, June, 2013.

Farzindar A., Inkpen D. (eds), *Proceedings of the Workshop on Semantic Analysis in Social Media*, Association for Computational Linguistics, April, 2012.

Farzindar A., Inkpen D., Gamon M., Nagarajan M. (eds), *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, Association for Computational Linguistics, April, 2014.

Farzindar A., Wael K., « A Survey of Techniques for Event Detection in Twitter », *Computational Intelligence*, vol. 31, n° 1, p. 132-164, 2015.

Gaio M., Sallaberry C., Nguyen V. T., « Typage de noms toponymiques à des fins d'indexation géographique », *Traitement Automatique des Langues*, vol. 53, n° 2, p. 143-176, 2012.

Hangya V., Berend G., Farkas R., « SZTE-NLP : Sentiment Detection on Twitter Messages », *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Association for Computational Linguistics, p. 549-553, 2013.

Hu M., Sun A., Lim E.-P., « Comments-oriented Blog Summarization by Sentence Extraction », *Proceedings of International Conference on Information and Knowledge Management (CIKM)*, ACM, p. 901-904, 2007.

Kong L., Schneider N., Swayamdipta S., Bhatia A., Dyer C., Smith N. A., « A Dependency Parser for Tweets », *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, p. 1001-1012, 2014.

Lin H., Bilmes J., Xie S., « Graph-based Submodular Selection for Extractive Summarization », *The eleventh biannual IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, IEEE, p. 381-386, 2009.

Melville P., Sindhwani V., Lawrence R. D., « Social Media Analytics : Channeling the Power of the Blogosphere for Marketing Insight », *Proceedings of the Workshop on Information in Networks (WIN)*, 2009.

Metzler D., Dumais S., Meek C., « Similarity Measures for Short Segments of Text », *Advances in Information Retrieval*, vol. 4425 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, p. 16-27, 2007.

Moncla L., Renteria-Agualimpia W., Nogueras-Iso J., Gaio M., « Geocoding for texts with fine-grain toponyms : an experiment on a geoparsed hiking descriptions corpus », *Proceedings*

of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, p. 183-192, 2014.

Moturu S., Quantifying the Trustworthiness of User-Generated Social Media Content, PhD thesis, Arizona State University, 2009.

Nallapati R., Feng A., Peng F., Allan J., « Event Threading Within News Topics », *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management (CIKM)*, ACM, p. 446-453, 2004.

Neviarouskaya A., Prendinger H., Ishizuka M., « Affect Analysis Model : Novel Rule-based Approach to Affect Sensing from Text », *Natural Language Engineering*, vol. 17, n° 1, p. 95-135, January, 2011.

Osborne M., Moran S., McCreadie R., Von Lunen A., Sykora M., Cano E., Ireson N., Macdonald C., Ounis I., He Y., Jackson T., Ciravegna F., O'Brien A., « Real-Time Detection, Tracking, and Monitoring of Automatically Discovered Events in Social Media », *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, Association for Computational Linguistics, p. 37-42, 2014.

Pang B., Lee L., Vaithyanathan S., « Thumbs Up ? : Sentiment Classification Using Machine Learning Techniques », *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 79-86, 2002.

Qadir A., Riloff E., « Learning Emotion Indicators from Tweets : Hashtags, Hashtag Patterns, and Phrases », *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, p. 1203-1209, 2014.

Sharifi B., Hutton M.-A., Kalita J. K., « Experiments in Microblog Summarization », *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, IEEE, p. 49-56, 2010.

Tang J., Chang Y., Liu H., « Mining Social Media with Social Theories : A Survey », *SIGKDD Explor. Newsl.*, vol. 15, n° 2, p. 20-29, June, 2014.

Turney P. D., « Thumbs Up or Thumbs Down ? : Semantic Orientation Applied to Unsupervised Classification of Reviews », *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, Association for Computational Linguistics, p. 417-424, 2002.

Zhou L., Hovy E. H., « On the Summarization of Dynamically Introduced Information : Online Discussions and Blogs. », *AAAI Spring Symposium : Computational Approaches to Analyzing Weblogs*, p. 237, 2006.

Détection d'évènements à partir de Twitter

Houssem Eddine DRIDI[1]* — **Guy LAPALME****

* *Druide informatique inc.*
1435 rue Saint-Alexandre, bureau 1040
Montréal, Québec, Canada H3A 2G4
houssemeddine.dridi@gmail.com
** *RALI - Département d'informatique et de recherche opérationnelle*
Université de Montréal
C.P. 6128, Succ Centre-Ville
Montréal, Québec, Canada H3C 3J7
lapalme@iro.umontreal.ca

RÉSUMÉ. *Nous présentons un système pour déterminer, à partir des données de Twitter, les évènements qui suscitent de l'intérêt d'utilisateurs au cours d'une période donnée ainsi que les dates saillantes de chaque évènement. Un évènement est représenté par plusieurs termes dont la fréquence augmente brusquement à un ou plusieurs moments durant la période analysée. Afin de déterminer les termes (notamment les hashtags) portant sur un même sujet, nous proposons des méthodes pour les regrouper: des méthodes phonétiques adaptées au mode d'écriture utilisé par les utilisateurs et des méthodes statistiques. Pour sélectionner l'ensemble des évènements, nous avons utilisé trois critères :* fréquence, variation *et* Tf·Idf.

ABSTRACT. *We present a system for finding, from Twitter data, events that raised the interest of users within a given time period and the important dates for each event. An event is represented by many terms whose frequency increases suddenly at one or more moments during the analysed period. In order to determine the terms (especially the hashtags) dealing with a topic, we propose methods to cluster similar terms: phonetic methods adapted to the writing mode used by users and some statistical methods. In order to select the set of events, we used three main criteria:* frequency, variation *and* Tf·Idf.

MOTS-CLÉS : *Twitter, hashtags, évènement, similarité sémantique,* DBscan.

KEYWORDS: *Twitter, hashtags, event, semantic similarity,* DBscan.

1. travail effectué au RALI-DIRO Université de Montréal

TAL. Volume 54 – n° 3/2013, pages 17 à 39

1. Introduction

Plusieurs recherches ont montré que les données publiées par les internautes sur les sites de médias sociaux, notamment Twitter, reflètent presque en temps réel l'intérêt du public. Twitter limite à 140 le nombre de caractères utilisés dans un message incluant possiblement des hyperliens. Dans un tweet, les sujets peuvent être étiquetés avec un mot *hashtag*, un mot précédé par un dièse `#` (*hash* en anglais). En cliquant sur un hashtag, la liste des tweets ayant le même hashtag s'affiche. Voici un exemple de tweet : `les manifestants se sont dispersés. #manifencours #ggi`. Les sujets sont `manifencours` et `ggi`. En cliquant sur `#ggi` la liste des tweets ayant comme sujet `ggi` s'affiche.

Comme tous les médias sociaux, les utilisateurs inscrits sont en mesure d'établir des relations entre eux, un utilisateur pouvant s'abonner à d'autres ce qui lui permet de consulter leurs messages au moment de sa connexion.

Le contenu d'un tweet peut être un avis, une information ou un témoignage. La vaste communauté de Twitter, le haut taux d'utilisation, plus de 500 millions de tweets par jour, et la variété des intérêts des utilisateurs accumulent des informations sur des évènements locaux (par exemple, *grève sur la hausse des frais de scolarité au Québec*) ou internationaux (*décès de Michael Jackson*). Comme nous le détaillons à la section 2, plusieurs études (Sutton *et al.*, 2008 ; Kwak *et al.*, 2010 ; Becker *et al.*, 2011 ; Jianshu et Bu-Sung, 2011 ; Ozdikis *et al.*, 2012a ; Ozdikis *et al.*, 2012b) ont montré que Twitter est une source intarissable pour dégager les intentions ou même les émotions des utilisateurs. Contrairement aux autres plates-formes de médias sociaux (*e.g. Facebook*), le contenu de Twitter est public et accessible *via* des interfaces de programmation. Tous ces facteurs nous ont encouragés à utiliser Twitter pour réaliser notre objectif, soit l'identification d'évènements qui stimulent l'intérêt des utilisateurs à un moment donné.

Nous considérons un évènement comme *quelque chose* qui arrive sur une seule journée et à un seul endroit, par exemple une manifestation, ou bien qui s'étend sur plusieurs jours ou plusieurs endroits, par exemple une épidémie. Un évènement sera représenté par un ensemble de *termes*[1] dont la fréquence augmente brusquement à un ou plusieurs moments durant la période analysée. Comme les hashtags permettent de donner une idée générale sur les sujets discutés dans un tweet, la majorité de nos méthodes utilisent ces éléments afin de déterminer les sujets saillants.

Nous avons expérimenté avec des tweets portant sur la Tunisie, la plupart écrits par des Tunisiens. Nous avons été amenés vers ce type de textes à cause de nos compétences qui nous permettent de comprendre le français et l'arabe, en particulier le mode d'écriture des Tunisiens, détaillé à la section 3, qui comporte des abréviations, des fautes de grammaire et d'orthographe, des mots arabes écrits avec des alphabets français et des chiffres et plusieurs langues à l'intérieur d'un même tweet. Il nous

1. Ici, selon le contexte, un terme peut correspondre à un mot, à un groupe de mots ou à un hashtag

était ainsi plus facile de déterminer la précision de notre système étant donné notre connaissance de l'actualité en Tunisie. Notre corpus est composé de 276 505 tweets collectés pendant 67 jours (du 8 février au 15 avril 2012).

À la différence de travaux portant sur la détection d'évènements à partir de documents longs et structurés tels que ceux analysés dans le cadre du projet TimeML (Pustejovsky *et al.*, 2010) ou du *Topic Detection and Tracking* (TDT) (Allan *et al.*, 1998), notre tâche est compliquée par la taille limitée et le type particulier d'écriture des tweets écrits en dialecte tunisien. La limite de taille des tweets reste toutefois un avantage car un utilisateur ne peut se disperser et ne traite donc que d'un seul sujet, contrairement à des textes plus longs où il peut être plus difficile de déterminer l'évènement relaté. Farzindar et Khreich (2013) présentent un panorama complet de la problématique de la détection d'évènements avec Twitter.

Intuitivement, l'augmentation brusque de la fréquence d'un terme devrait indiquer la présence d'un sujet saillant ou évènement. Pour le vérifier, nous avons calculé les fréquences de différents termes, notamment les hashtags. Effectivement, nous avons constaté que les fréquences de certains termes avaient tendance à augmenter brusquement lors d'un évènement. Si chaque terme se référait à un sujet distinct, nous pourrions distinguer facilement les évènements. Cependant, un sujet est souvent représenté par plus d'un terme. Par exemple, la disparition de l'avion *Boeing 777* du *vol MH370* de la *Malaysia Airlines* le 8 mars 2014, a provoqué l'apparition de plusieurs hashtags référant à cet évènement : *#PrayForMH370*, *#MH370*, *#MH370Flight*, *#MalaysiaAirlines*...

Il ne suffit donc pas de calculer la fréquence de chaque terme séparément, il faut plutôt les regrouper quand ils réfèrent au même sujet, pour ensuite calculer la fréquence de chaque cluster afin de déterminer l'évènement le plus important. Certains termes (hashtags) du même cluster pouvant avoir été créés avant d'autres, le regroupement des termes sert aussi à déterminer la durée de l'évènement.

La figure 1 montre les différentes expériences que nous avons menées et qui sont décrites dans les sections suivantes.

La tâche initiale était d'extraire, d'une façon continue, les tweets à partir de Twitter. La méthode d'extraction du corpus et ses caractéristiques sont décrites à la section 3.

Notre première tâche est donc de regrouper les termes référant à un même sujet. Pour ce faire, nous avons développé trois méthodes[2].

Normalisation des hashtags ① Les utilisateurs des médias sociaux commettent souvent des fautes d'orthographe. Dans notre cas, ce phénomène est amplifié par le fait que les Tunisiens ont tendance à écrire des mots arabes en utilisant l'alphabet latin et des chiffres, chaque utilisateur translittérant le mot de sa façon. Pour normaliser ces hashtags, nous avons eu recours à des algorithmes phonétiques de type *Soundex* pour supporter le dialecte tunisien, afin d'attribuer le même

2. Les nombres encerclés font référence à ceux de la figure 1

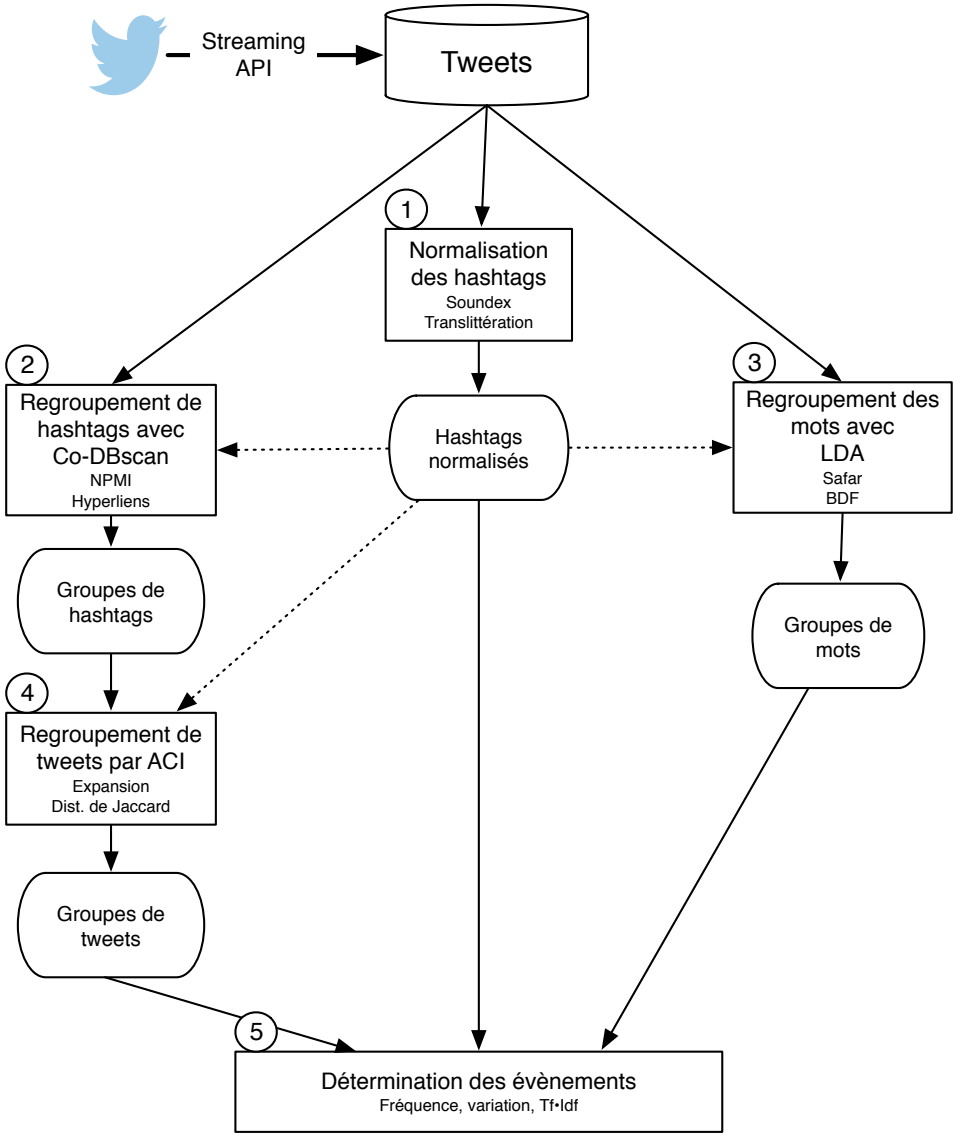

Figure 1. *Organisation des expériences menées avec le système de détection d'évènements : un rectangle représente un traitement identifié par un numéro encerclé détaillé dans le texte ; on indique en italique la méthode impliquée à cette étape. Un rectangle arrondi représente un résultat obtenu après traitement.*

code aux hashtags avec une prononciation semblable. Nous proposons, également, un algorithme de translittération qui permet de coder les hashtags écrits en alphabet arabe avec le même code que leurs semblables écrits en alphabet latin. La section 4.1 détaille ce processus.

Regroupement de hashtags avec $CoDBscan$ ② Nous avons développé une variante de l'algorithme de regroupement *DBscan* qui tient compte de la distance entre les éléments pour regrouper les hashtags similaires. Le calcul de la distance entre les hashtags est fondé sur leur cooccurrence et celle des hyperliens. Nous avons profité également des résultats obtenus par la tâche de normalisation pour améliorer le regroupement. Plus de détails dans la section 4.2.

Regroupement des mots similaires ③ Nous avons utilisé LDA (*Latent Dirichlet Allocation*), une technique statistique pour détecter les sujets d'une collection de documents. LDA considère un document comme un mélange de sujets latents, un sujet étant lui-même représenté comme une distribution de mots ayant tendance à apparaître ensemble. Nous avons profité de cette distribution pour établir les termes d'un même sujet. La section 4.3 explique ce processus.

Regroupement de tweets par un algorithme de clustering incrémental ④ Nous regroupons les tweets similaires en comparant les hashtags trouvés dans les tweets avec la mesure de Jaccard. Dans ce processus, nous avons étendu les tweets par d'autres hashtags similaires, en utilisant des clusters de hashtags, obtenus par des variantes de $CoDBscan$, afin d'améliorer le regroupement. Voir la section 4.4 pour plus de détails.

Détermination des évènements ⑤ Une fois ces regroupements effectués, nous avons utilisé les clusters de hashtags et de mots obtenus pour étiqueter chaque tweet avec un sujet. Nous supposons qu'un tweet porte sur un sujet si au moins un de ses termes est présent. Cette tâche a permis de calculer la fréquence quotidienne de chaque sujet. Nous identifions l'ensemble de sujets référant à des évènements en fonction de trois critères : fréquence, variation (ou écart-type) et *Tf·Idf* pour lesquels des sujets correspondent à des pics. Cette méthodologie est présentée à la section 5.

Nos résultats (section 6) ont été évalués par des personnes familières avec les évènements qui se sont déroulés en Tunisie à l'aide d'une application Web que nous avons développée pour en faciliter l'annotation. La section 7 présente des extensions pour des travaux futurs et la section 8 conclut en rappelant nos principales contributions.

2. Travaux antérieurs en détection des évènements

Les microblogs sont un excellent moyen pour diffuser des informations, discuter des évènements et y donner des avis. Kwak *et al.* (2010) ont constaté que les utilisateurs de Twitter diffusent parfois des nouvelles avant les journaux, la télévision ou la

radio. Sutton *et al.* (2008), dans une étude sur les incendies des forêts en Californie en 2007, ont montré que Twitter a représenté une source d'information importante pour les citoyens ; ils ont même constaté que les médias traditionnels se sont tournés vers Twitter pour obtenir des renseignements.

Les recherches qui s'intéressent à la détection des évènements à partir des articles journalistiques sont fondées sur des techniques de traitement automatique de la langue naturelle (Makkonen *et al.*, 2004 ; Zhang *et al.*, 2007) dont l'extraction d'entités nommées. Toutefois, l'application de ces techniques sur les tweets est plus difficile compte tenu de leur faible quantité d'informations. Plusieurs recherches ont montré que le contenu de ces outils (notamment Twitter) reflète bien l'intérêt et les préoccupations des utilisateurs en temps réel.

Jianshu et Bu-Sung (2011) ont proposé une méthode fondée sur la fréquence quotidienne des termes dans le corpus. La fréquence de chaque terme est représentée sous forme d'un signal. Ces signaux sont analysés par ondelettes pour déterminer quand et comment la fréquence du signal change dans le temps (Kaiser, 2011). Les auteurs ont considéré que les termes avec des signaux similaires représentaient le même évènement, la similarité entre signaux étant calculée par corrélation croisée.

Ozdikis *et al.* (2012a) et Ozdikis *et al.* (2012b) ont effectué une expansion sémantique des termes présentés dans les tweets qui s'appuie sur la cooccurrence des termes afin de regrouper les tweets selon leur similarité. Chaque tweet est représenté sous forme d'un vecteur de termes. La similarité entre deux tweets est calculée par cosinus de similarité. Les auteurs ont considéré que chaque cluster de tweets représente un évènement. Ce travail a montré que les hashtags sont de bons indicateurs pour détecter les évènements à partir des tweets. En outre, l'expansion sémantique augmente le nombre de tweets portant sur un même évènement ce qui augmente également son importance et allonge la durée d'un évènement.

Becker *et al.* (2011) ont proposé une approche pour identifier, parmi tous les tweets, ceux qui décrivent des évènements. Plusieurs tweets sont tout simplement une conversation entre amis ou des opinions. Les auteurs ont implémenté un algorithme de regroupement en ligne, qui affecte chaque nouveau tweet à un cluster adéquat. Par la suite, ils ont appliqué un algorithme de classification pour déterminer si un cluster (selon le contenu de ses tweets associés) porte sur un évènement ou non.

Comme les autres tâches, la détection d'évènements s'appuie sur les termes pour déterminer les sujets saillants survenus durant une période donnée. Toutefois, pour la détection d'évènements non connus *a priori*, il est difficile d'utiliser un ensemble prédéfini pour déterminer des évènements qu'on ne connaît pas. Les travaux qui s'intéressent à cette tâche cherchent à regrouper les tweets similaires puis à déterminer parmi les clusters obtenus ceux qui réfèrent à des évènements. La similarité entre les tweets s'appuie souvent sur leur contenu textuel. Nous utilisons une approche similaire, mais nous suggérons une méthode plus simple fondée sur la fréquence des clusters de termes. Malgré sa simplicité, notre méthode repère les sujets saillants au

cours d'une période. Notre système détecte non seulement les évènements, mais il en détermine les dates saillantes.

3. Description du corpus

3.1. *Caractéristiques des données*

Nous avons extrait les tweets d'une façon continue pendant plusieurs jours en utilisant la *streaming API*[3]. Pour recueillir des tweets, nous avons utilisé l'ensemble de mots-clés décrit dans le tableau 3.1. Nous avons récupéré 276 505 tweets entre le 8 février 2012 et le 15 avril 2012. Une fois enlevés les hashtags, les utilisateurs mentionnés et les hyperliens, les tweets contiennent en moyenne 12,4 mots, le tweet le plus long en contenant 37. Le tableau 3.1 présente quelques statistiques sur le corpus.

Mots-clés	**Définition**
`marzouki`	Président actuel de la Tunisie
`hammadi jebali`	Premier ministre dans cette période
`Tunisie`, `tounes`, `Tunisia`	`tounes` est la prononciation arabe de Tunisie
`tnelec`	Les élections tunisiennes
`sebsi`	Ex-Premier ministre après la révolution tunisienne
`nahdha`, `ennahdha`	Le parti au pouvoir durant cette période
`ghannouchi`	Chef d'`ennahda`
`sidi bouzid`	La région où la révolution tunisienne a commencé
`14jan`	14 janvier, date de fuite du président déchu Ben Ali

Tableau 1. *Mots-clés utilisés pour extraire notre corpus*

Numbre de tweets	276 505
Nombre des retweets	32 890
Nombre d'utilisateurs distincts	26 093
Nombre de tweets qui contiennent au moins un hashtag	147 395
Nombre de tweets qui contiennent au moins un utilisateur mentionné	88 595
Nombre de tweets qui contiennent au moins un hyperlien	168 309
Nombre de hashtags distincts	12 218

Tableau 2. *Statistiques sur les tweets de notre corpus*

3. `https://dev.twitter.com/streaming/overview`

3.2. *Dialecte tunisien*

Le dialecte tunisien est la langue employée par tous les Tunisiens, appelé *darija*. Il diffère de l'arabe standard, il est très influencé par la langue française, mais il intègre parfois des mots d'autres langues comme l'anglais, le punique, le berbère ou l'italien.

Le mode d'écriture employé par les Tunisiens dans les SMS et les médias sociaux présente d'autres caractéristiques : un même mot peut être écrit en alphabet latin ou arabe et parfois en mélangeant les deux alphabets. Lorsqu'un mot arabe est écrit en alphabet latin, les lettres arabes ne pouvant être transcrites directement sont remplacées par un chiffre dont la forme rappelle vaguement la lettre en arabe ou avec deux lettres qui rappellent la prononciation de la lettre. Le même mot peut donc être écrit de plusieurs façons, d'où la nécessité d'une certaine normalisation. Dridi (2014) présente plusieurs exemples de ce type d'écriture.

Certains travaux se sont intéressés au dialecte tunisien (Boujelbane, 2013), mais ils traitaient des textes plus longs en arabe plutôt que des tweets très courts combinant souvent les alphabets latin et arabe. Nous n'avons pas trouvé de ressources linguistiques permettant de déterminer les relations sémantiques entre des termes du dialecte tunisien, et même si nous en avions eues, nous aurions toujours eu besoin de déterminer les termes sémantiquement similaires à cause de l'évolution dynamique du vocabulaire dans les médias sociaux.

4. Regroupement

Nos méthodes s'appuient sur les termes trouvés dans les tweets pour déterminer les sujets saillants ; or, un sujet est souvent représenté par plus d'un terme. Il est donc important de regrouper les termes référant un même sujet, sinon chaque terme dans le corpus représentera un sujet différent.

Généralement, les textes générés par les utilisateurs sur le Web, notamment dans les microblogs, contiennent des mots non standard, car les utilisateurs commettent souvent des fautes d'orthographe et utilisent des abréviations produisant ainsi plusieurs variantes pour un même terme. Plusieurs travaux (Clark et Araki, 2011 ; Sproat *et al.*, 2001) proposent de normaliser automatiquement les termes, c'est-à-dire transformer toutes les variantes en un terme unique. Cette section présente notre approche à ce problème après avoir décrit quelques particularités de notre corpus.

4.1. *Normalisation des hashtags*

Dans cette section, nous présentons les techniques que nous avons appliquées pour normaliser les termes (① dans la figure 1). Les utilisateurs de microblogs commettent souvent des fautes d'orthographe créant ainsi plusieurs variantes pour un même terme. Ce problème a été déjà traité par les systèmes de correction orthographique à l'aide d'algorithmes phonétiques. Ces algorithmes indexent les mots selon leur prononcia-

tion. Le principe consiste à utiliser la prononciation d'un mot mal écrit pour prédire le bon mot, avec la même prononciation, qui lui correspond. Nous avons utilisé l'algorithme de *Soundex* (Russell, 1918) afin de normaliser les termes qui ont une prononciation similaire.

Comme les prononciations varient d'une langue à une autre, il existe plusieurs variantes de *Soundex*. Nous avons appliqué un *Soundex* pour le français standard puisque la prononciation des Tunisiens ressemble à celle des Français et nous l'avons adapté au dialecte tunisien. Nous avons également proposé un algorithme de translittération afin de regrouper les hashtags écrits en alphabet arabe avec leur équivalent en alphabet latin. Toutefois, les utilisateurs utilisent souvent des hashtags dérivés des dates pour référer à des évènements. Comme le *Soundex* n'est pas adapté à ce type de données, nous avons traité ces hashtags de façon particulière avec des expressions régulières. Dridi (2014) détaille ce processus.

4.2. *Regroupement des hashtags*

La relation sémantique entre deux termes sert à déterminer leurs degrés d'association. Cette information joue un rôle important dans plusieurs domaines du TAL tels que la construction automatique des thesaurus, la recherche d'information... Par exemple, il est utile d'utiliser les termes similaires à ceux spécifiés dans la requête de l'utilisateur pour récupérer les documents pertinents. Plusieurs travaux se sont appuyés sur des bases construites manuellement par des linguistes (*e.g.* WordNet) pour déterminer la relation sémantique entre les termes. Ces bases contiennent des informations indiquant le type de relation (synonyme, antonyme, hyperonyme...) entre les termes. Cependant, elles ne couvrent pas les dialectes ni le mode d'écriture (fautes, abréviations) employés dans les médias sociaux. En outre, le vocabulaire employé dans les médias sociaux est enrichi fréquemment par de nouveaux termes inventés par les utilisateurs (par exemple *produits*, *personnes*, *parti politique*...).

Une autre approche, à base de techniques statistiques, permet de fournir une information quantitative indiquant le degré de la similarité sémantique entre les termes. Cette information est estimée à partir des données observées, en se fondant sur la notion de cooccurrence, soit l'apparition simultanée de deux ou plusieurs termes dans une même fenêtre. Une fenêtre peut être un paragraphe ou une phrase. Deux termes qui cooccurrent appartiennent souvent à un même contexte. Par exemple, les termes *loi* et *avocat* apparaissent fréquemment ensemble dans un même contexte : *la justice*.

Nous avons implémenté des méthodes s'appuyant sur la cooccurrence (② dans la figure 1) afin de regrouper les hashtags d'un sujet commun. Étant donné que les tweets sont courts, nous avons considéré le tweet entier comme fenêtre. Nous avons constaté que les hashtags qui cooccurrent fréquemment sont similaires ou réfèrent à un même sujet.

Pour mesurer le degré de relation entre deux hashtags, nous avons utilisé la mesure *Pointwise Mutual Information* (*PMI*) (Church et Hanks, 1989). *PMI* mesure la quan-

tité d'informations apportée pour la présence simultanée d'une paire de termes, dans notre cas les hashtags H_i.

$$PMI(H_i, H_j) = \log(\frac{P(H_i\&H_j)}{P(H_i)P(H_j)}) = \log(\frac{N * a}{(a+b) * (a+c)})$$

$P(H_i\&H_j)$ est la probabilité que H_i et H_j apparaissent ensemble dans un même tweet. $P(H_i)P(H_j)$ est la probabilité que H_i et H_j apparaissent ensemble, s'ils sont statistiquement indépendants. Le ratio $P(H_i\&H_j)$ et $P(H_i)P(H_j)$ mesure le degré de dépendance entre H_i et H_j. ***PMI*** est maximisé lorsque H_i et H_j sont parfaitement associés. *N* est le nombre de tweets considérés ; a est le nombre de fois que H_i et H_j apparaissent ensemble, b le nombre de fois que H_i est présent, mais que H_j est absent alors que c est le nombre de fois que H_j est présent, mais où H_i est absent.

Pour la plupart des algorithmes de regroupement, l'utilisateur doit disposer de suffisamment de connaissances sur les données pour déterminer le nombre de clusters. À titre d'exemple, l'algorithme *k-moyenne* nécessite de spécifier à l'avance le nombre *k* de clusters à utiliser. Comme il est difficile de déterminer la bonne la valeur pour *k*, il faut en tester plusieurs.

Dans notre cas, la détermination *a priori* du nombre de clusters n'est guère envisageable, puisque chaque cluster représente un sujet dont nous ne connaissons pas le nombre dans le corpus. En outre, le *k-moyenne* est incapable de gérer les bruits et les exceptions, car chaque objet doit être associé à un cluster. D'autres algorithmes de regroupement n'exigent pas de spécifier à l'avance le nombre de clusters, par exemple l'algorithme *Density-Based Spatial Clustering of Applications with Noise* (*DBscan*) qui s'appuie sur la notion de densité (Ester *et al.*, 1996).

DBscan nécessite un seuil ϵ. Comme l'intervalle des valeurs de *PMI* est inconnu puisqu'il dépend de caractéristiques du corpus, ceci complique le choix de ϵ. Nous avons donc normalisé les valeurs de *PMI* en utilisant la méthode proposée par Bouma (2009).

$$\textit{NPMI}(e_i, e_j) = \log(\frac{P(e_i\&e_j)}{P(e_i)P(e_j)})/ - \log P(e_i\&e_j) = \textit{PMI}/ - \log P(e_i\&e_j)$$

P(e_i&e_j) est la probabilité que e_i et e_j apparaissent ensemble. La valeur de *NPMI* est donc comprise dans l'intervalle $[-1, 1]$. *NPMI*(e_i, e_j) vaut 1 lorsque e_i et e_j sont entièrement dépendants et -1 s'ils sont complètement indépendants. Cet intervalle fermé facilite la détermination de ϵ.

Nous avons appliqué *DBscan* pour regrouper les hashtags similaires. Nous avons considéré que les hashtags, qui se trouvent dans un même cluster, représentent un même évènement. Cependant, nous avons constaté que *DBscan* regroupe des hashtags qui ne sont pas vraiment similaires. Ceci est dû à un phénomène de transition lors de la fusion. Par exemple, le hashtag *#manifestation* pourrait appartenir à deux voisinages

$C \leftarrow 0$ *// initialiser le nombre de clusters à 0*
for chaque hashtag *h* non visité **do**
  ϵ-$voisinage \leftarrow$ epsilonVoisinage(*h*, ϵ)
  if tailleDe(ϵ-$voisinage$) <*MinHstgs* **then**
    marquer *h* comme $NOISE$ *// h n'appartient à aucun cluster*
  else {*il y a au moins* MinHstgs *points voisins à* h}
    $C \leftarrow C + 1$
    ajouter *h* au cluster *C*
    for chaque hashtag *h'* de ϵ-$voisinage$ **do**
      ϵ-$voisinage' \leftarrow$ hashtags de *H* de valeur *PMI* supérieure ou égale à ϵ
      if tailleDe(ϵ-$voisinage'$) $\geq$ *MinHstgs* **then**
        ϵ-$voisinage \leftarrow \epsilon$-$voisinage \cup \epsilon$-$voisinage'$
      if $Cohesion(C, h') > \lambda$ **then**
        ajouter *h'* au cluster *C*

Figure 2. *L'algorithme $CoDBscan$ où ϵ est la valeur minimale de similarité entre deux hashtags),* MinHstgs *est le nombre minimal de hashtags dans ϵ-voisinage) et* H *est l'ensemble de hashtags. λ est fixé à 0,1.*

représentant deux évènements différents, pendant lesquels des manifestations ont été organisées. L'application de la fusion regrouperait des hashtags portant sur deux sujets différents. Afin d'améliorer la cohésion à l'intérieur d'un cluster, nous ne les fusionnons pas directement, mais nous vérifions le degré de similarité de chaque élément d'un cluster avec tous les éléments d'un autre avant de les regrouper. Cette variante de *DBscan*, notée $CoDBscan$ pour *CohesiveDBScan*, est présentée à la figure 2.

Pour déterminer le degré de similarité, nous utilisons la fonction $Cohesion(C, e)$ déterminée comme suit :

$$Cohesion(C, e) = \frac{1}{|C|} \sum_{i=1}^{|C|} Similarity(h_i, e) \quad [1]$$

e est un hashtag dont on va mesurer le degré de similarité avec les hashtags trouvés dans un cluster *C*. *Cohesion(C,e)* est la moyenne des valeurs de *PMI* entre le hashtag *e* et ceux trouvés dans *C*. Plus la valeur de *Cohesion(C,e)* est grande plus le hashtag *e* est similaire aux hashtags trouvés dans *C*. Pour ajouter *e* à *C* la valeur de *Cohesion(C,e)* doit dépasser un certain seuil λ défini manuellement, 0,1 dans nos expériences. Une illustration du fonctionnement de cet algorithme est donné dans (Dridi, 2014).

Étant donné la taille réduite d'un tweet, les utilisateurs ne peuvent pas décrire des évènements, échanger des informations ou exprimer leur avis d'une manière efficace. Pour contourner cette limite, Twitter offre à ses utilisateurs la possibilité d'ajouter des hyperliens vers des pages externes permettant de mieux expliquer et détailler le contenu d'un tweet. Un hyperlien peut désigner une page qui contient un texte, des

images, de l'audio ou de la vidéo. Dans notre corpus, 62 % des tweets contiennent au moins un hyperlien qui constitue une information intéressante pour déterminer le degré d'association entre les termes, notamment les hashtags. Certaines URL sont trop longues et dépassent la taille permise pour un tweet, c'est pourquoi l'interface de Twitter raccourcit automatiquement les URL avec un service de réduction d'URL. Comme un tel service génère toujours la même URL pour la même entrée, même tapée par des utilisateurs différents, nous avons comparé les URL courtes plutôt que les URL finales.

Nous avons développé une autre méthode de regroupement à base d'hyperliens : nous avons supposé que deux hashtags qui apparaissent avec les mêmes hyperliens représentent le ou les mêmes sujets détaillés par ces hyperliens. Pour regrouper les hashtags, nous avons utilisé l'algorithme $CoDBscan$ de la figure 2 en ajustant la mesure de similarité entre deux hashtags pour tenir compte du fait qu'ils cooccurrent avec les mêmes liens, cette méthode est notée $CoDBscan_{hyper}$.

4.3. *Regroupement des mots*

Afin de regrouper des mots dans les tweets (③ dans la figure 1), nous avons considéré la technique du *Topic Model (TM)* qui identifie des sujets abstraits à partir d'une collection de documents. *TM* considère que chaque document peut être représenté comme un mélange de sujets latents, où un sujet est lui-même représenté comme une distribution de mots ayant tendance à cooccurrer, les mots fortement liés à un sujet ayant des valeurs plus grandes. Nous avons profité de cette distribution pour obtenir les termes qui portent sur un même sujet. Les algorithmes de *TM* utilisent la modélisation de sac de mots qui représentent chaque document par les fréquences des mots qui le composent. Cela permet d'ignorer la syntaxe des phrases pour se concentrer sur les termes trouvés dans le document. Le nombre de sujets est un paramètre qui doit être donné avant l'application de *TM*. Dans ce travail, nous avons appliqué l'algorithme *Latent Dirichlet Allocation* (LDA) (Blei *et al.*, 2003) à l'aide de la librairie *Mallet* (McCallum, 2002). Nous avons considéré chaque tweet comme étant un document différent. Après la tokenization des tweets, nous avons appliqué des prétraitements : nous avons supprimé les *usernames* (identifiant d'un utilisateur dans Twitter) et les hyperliens ; nous avons supprimé les mots-clés utilisés pour extraire les tweets ; nous avons éliminé les mots vides anglais, français, arabes et leurs équivalents en dialecte tunisien.

Pour obtenir des données plus riches, nous avons regroupé au sein d'un seul document les tweets partageant un même hashtag, même s'ils sont en différentes langues (*e.g.* en arabe ou en français). Chaque cluster va représenter un document, sous forme de plusieurs tweets. Avec cette technique, les termes qui sont sémantiquement proches seront regroupés, ce qui permet d'obtenir un vocabulaire beaucoup plus riche qu'en se limitant à un simple tweet. Cette méthode amène LDA à retourner des sujets plus significatifs quand il est utilisé sur un long document que sur un court tweet. LDA est toutefois incapable d'identifier les sens des termes, par exemple *parlent* et *parle* sont

considérés comme différents même s'ils correspondent au même verbe. Nos tweets étant écrits en français et en arabe (incluant le dialecte tunisien), nous avons utilisé *BDF*, un lemmatiseur français de notre laboratoire, pour remplacer un mot en français par son lemme alors que pour les mots écrits en alphabet arabe nous avons appliqué le *stemmer* de la librairie *Safar* [4].

4.4. *Regroupement des tweets*

Les mesures de similarité s'appuient sur les hashtags qui sont en commun dans deux tweets. Cependant, les termes, sémantiquement similaires ou écrits dans une forme différente (fautes d'orthographe, abréviation...), ne sont pas considérés comme des termes en commun. Par exemple, si un tweet contient le terme `Ghnem` et un autre contient le terme `Ghanim`, ces deux termes ne seront pas considérés comme des termes en commun même s'ils sont similaires.

Pour améliorer les résultats obtenus par les mesures de similarité afin de détecter les termes en commun entre deux documents même s'ils sont écrits d'une manière différente, nous avons décidé d'étendre les tweets par des hashtags sémantiquement similaires. Cette technique, couramment en recherche d'information afin d'améliorer les résultats de recherche, étend la requête avec d'autres termes reliés. Certaines méthodes utilisent des ressources externes telles que des thesaurus (par exemple WordNet) pour enrichir la requête.

Nous adoptons le même principe en enrichissant les tweets par d'autres hashtags. Cependant, le recours à des ressources externes prédéfinies est difficile étant donné le peu de ressources pour le dialecte tunisien. Un thesaurus contient un ensemble fini de termes, mais le vocabulaire utilisé dans les médias sociaux est évolutif car les internautes produisent souvent de nouveaux termes (abréviations, conventions, etc.) à mesure qu'apparaissent de nouvelles personnes et technologies. Pour étendre les tweets, nous avons exploité les clusters de hashtags sémantiquement similaires obtenus par les méthodes de regroupement et de normalisation présentées à la section précédente. Chaque hashtag est remplacé, le cas échéant, par le cluster de hashtags auquel il appartient. Si un hashtag n'appartient à aucun cluster, nous utilisons le hashtag lui-même.

Les travaux sur la détection des évènements utilisent des techniques de regroupement afin de répartir le corpus en clusters dont les documents sont supposés discuter sur le même sujet (évènement). Les méthodes de regroupement existantes représentent un document sous forme d'un vecteur de traits ainsi que l'importance de chaque trait dans le document. Les traits sont généralement les termes du vocabulaire utilisé dans le corpus. Dans cette représentation vectorielle, l'importance d'un terme dans un document est donnée par le produit *Term Frequency-Inverse Document Frequency* (*Tf·Idf*).

Les algorithmes de regroupement s'appuient sur ces mesures de similarité afin de comparer chaque paire de documents. Étant donné la richesse du vocabulaire de notre

4. `http://sibawayh.emi.ac.ma/safar/index.php`

corpus due principalement à la variété d'écriture de plusieurs termes et aux différentes langues utilisées, le nombre de traits utilisés pour représenter un document (tweet) est immense. Un tweet contenant peu de termes, dans notre corpus environ 7, la plupart des traits d'un tweet sont donc à 0. Nous avons constaté que la plupart des tweets ne contiennent pas de termes qui se répètent. En appliquant *Tf·Idf*, la valeur *Tf* vaut souvent 1 pour les termes présents dans les tweets. Étant donné ces deux caractéristiques, nous avons décidé de représenter les tweets par des vecteurs binaires en considérant seulement les termes, dont la valeur est plus grande ou égale à 1, et nous utilisons l'indice de Jaccard pour comparer une paire de tweets. La mesure du cosinus de similarité pourrait être plus efficace sur des grands textes, mais nous n'avons pas besoin d'utiliser les poids (voir *supra*), l'indice de Jaccard est plus simple et donne de très bons résultats.

Un algorithme de clustering incrémental (④ dans la figure 1) est appliqué pour regrouper les tweets similaires sans fixer *a priori* le nombre de clusters à obtenir.

Le cluster C^* auquel est affecté un tweet tw_i est déterminé par

$$C^* = \arg\max_{C_j \in C} \frac{1}{|C_j|} \sum_{tw_k \in C_j} Similarity(tw_i, tw_k)$$

où

– C_j est le $j^{\text{ième}}$ cluster déjà créé ;

– $Similarity(tw_i, tw_k)$ calcule la similarité, entre tw_i et un tweet tw_k de C_j, en utilisant l'indice de Jaccard ;

– $|C_j|$ est le nombre de tweets dans C_j.

Pour chaque cluster C_j existant, nous comparons tous ses tweets avec tw_i. C_j est considéré comme candidat pour tw_i si la moyenne de la similarité ($\frac{1}{|C_j|} \sum_{tw_k \in C_j} Similarity(tw_i, tw_k)$) de tw_i et des tweets de C_j dépasse un seuil ϵ que, suite à nos expériences préliminaires, nous avons fixé à 0,5.

Le candidat, qui a la valeur $\frac{1}{|C_j|} \sum_{tw \in C_j} Similarity(tw_i, tw)$ la plus élevée, est le cluster auquel tw_i va appartenir. Si l'ensemble de candidats est vide, nous créons un nouveau cluster contenant tw_i. Il n'est pas indispensable d'affecter tous les tweets à des clusters.

Nous obtenons ainsi des tweets avec plus d'informations ce qui permet aux mesures de similarité de mieux détecter les termes en commun entre une paire de tweets. Chaque cluster ne devrait contenir que des tweets discutant du même sujet, mais possiblement écrits à des dates différentes.

5. Détermination des évènements

Allan *et al.* (1998) ont défini un évènement par quelque chose d'unique qui arrive à un certain point dans le temps. Kleinberg (2003) fait remarquer qu'une tendance

ou un évènement qui suscitent de l'intérêt dans un flux de documents sont signalés par un regain d'activité signalé par une augmentation marquée de la fréquence de certains termes associés à l'évènement en question. La détermination des évènements (⑤ dans la figure 1) est un champ de recherche étudié depuis des années, il est souvent appelé *Topic Detection and Tracking* (TDT). La TDT est facilitée par l'existence de flux quotidiens des journaux sur le web. La motivation de cette tâche est l'implémentation d'un système d'alerte qui permet de détecter et d'analyser, à partir d'un flux de documents, les évènements majeurs (Allan, 2002). Les recherches, qui ont abordé ce thème, ont utilisé principalement des techniques de TAL (*e.g.* lemmatisation, détermination des parties du discours...) qui sont efficaces pour des documents contenant des informations bien structurées.

Nous pouvons distinguer deux types de travaux qui s'intéressent à l'identification des évènements.

Évènements connus *a priori* Ces travaux se concentrent sur les évènements dont les caractéristiques (type, nom, emplacement...) sont connues au préalable. Certains travaux (Sakaki *et al.*, 2010) s'intéressent à l'identification de documents (messages) qui discutent d'un évènement particulier (*e.g.* tremblement de terre, concert...) en formulant des requêtes contenant ses caractéristiques. Chakrabarti et Punera (2011) et Shamma *et al.* (2010) s'intéressent à la génération des résumés des messages qui discutent d'un évènement particulier. Petrovic *et al.* (2010) ont essayé de trouver le premier message qui discute d'un évènement précis.

Évènements inconnus Dans ce cas, les recherches s'intéressent à la détection des tendances en identifiant les sujets inédits, ou en croissance rapide, au sein d'un flux de documents (Kontostathis *et al.*, 2004). Plusieurs travaux ont montré que les utilisateurs de Twitter discutent et partagent des nouvelles à propos d'évènements imprévus (*e.g.* tremblement de terre). Pour cette raison les travaux qui s'intéressent à la détection des tendances (ou des évènements inconnus) ont eu recours à d'autres indices permettant de signaler la présence d'un évènement dans une période.

Un évènement *e*, au cours d'une période *T*, est représenté par un ensemble de traits F_e dont les fréquences augmentent brusquement à un ou plusieurs points t_e inclus dans *T*. Dans nos expériences, nous nous appuyons sur des fréquences quotidiennes.

Twitter offre déjà un service qui affiche à tout moment les dix tendances les plus fortes (sous forme de termes) dans la page d'accueil d'un utilisateur. Les tendances sont, par défaut, personnalisées par l'emplacement de l'utilisateur et changent régulièrement à un intervalle de quelques minutes. L'algorithme utilisé par ce service n'est pas diffusé, mais nous avons constaté que les tendances affichées semblent référer aux termes les plus fréquents à un moment donné. Les termes peuvent être des hashtags ou des mots dans les tweets. Ce service ne regroupe pas les termes qui représentent la même tendance, par exemple, lors du décès de Michael Jackson, la plupart des

tendances étaient autour de ce sujet : *Michael Jackson*, *MJ*, *King of Pop*... (Kwak *et al.*, 2010). Cependant, Twitter n'affiche pas les tendances pour toutes les régions, nous aurions aimé afficher les tendances pour la Tunisie, mais elle ne fait pas partie de la liste des régions géolocalisées par Twitter.

Étant donné le nombre important de tweets qui contiennent au moins un hashtag et la difficulté d'analyser les tweets, nous avons d'abord essayé de nous appuyer sur les hashtags pour identifier les sujets les plus discutés par les internautes. Nous avons constaté qu'en effet les hashtags donnaient une bonne idée des préoccupations des utilisateurs. Comme un sujet est souvent représenté par plus d'un hashtag, il est nécessaire d'identifier les hashtags sémantiquement similaires, c'est-à-dire discutant du même sujet, sinon chaque hashtag représentera un évènement différent.

Au début, nous avons calculé les fréquences de différents termes (hashtags et mots) trouvés dans notre corpus ; or nous avons constaté que cette méthode n'est pas très efficace parce qu'un évènement peut être référé par plus d'un terme. Pour cette raison, au lieu de calculer la fréquence quotidienne de chaque terme, il est préférable de calculer la fréquence quotidienne des termes d'un cluster représentant un évènement. Nous avons simplement incrémenté la fréquence quotidienne F_{gj} d'un cluster *g* au jour *j*, si au moins l'un des termes de g se trouve dans TW_{ij}, où TW_{ij} est le tweet i au jour j. Nous obtenons ainsi une fréquence quotidienne pour chaque évènement représenté par son cluster de termes.

Afin de détecter les évènements majeurs et de déterminer les dates où ces évènements ont eu lieu, nous avons utilisé la méthode proposée par Palshikar (2009) permettant de détecter les dates saillantes. Cette méthode détecte les pics dans une série temporelle. Elle prend comme entrée une série $f(s)$ et retourne les indices I qui correspondent aux pics. Dans notre cas, les indices sont les jours et les fréquences correspondent au nombre d'éléments d'un cluster pour cette journée.

Soit S la fonction qui permet de détecter les pics dans une période donnée :

$$S_i(k, f(s)) = \frac{\max\limits_{1\leq j\leq k}(f_i(s) - f_{i-j}(s)) + \max\limits_{1\leq j\leq k}(f_i(s) - f_{i+j}(s))}{2}$$

où *k* est un entier positif indiquant le nombre de voisins à considérer autour de chaque point $f_i(s)$ dans τ, les valeurs les plus appropriées de *k* étant comprises entre 3 et 5. Pour $f_i(s) \in f(s)$, S_i calcule la moyenne de la différence maximale entre les valeurs de *k* voisins à gauche et à droite de $i^{\text{ème}}$ élément. $f_i(s)$ est considéré comme un pic si :

$$S_i > 0 \textbf{ et } (S_i - mean) > stdv$$

où *mean* et *stdv* sont respectivement la moyenne et l'écart-type des valeurs positives de S_i. Dans nos expériences, *k* a été fixé à 3.

Pour déterminer les évènements saillants, nous avons testé trois critères : la fréquence quotidienne, la variation définie par l'écart-type de la fréquence quotidienne et une mesure inspirée du *Tf·Idf* où :

– Tf : correspond à la fréquence quotidienne du sujet S durant la période ;

– $Idf = \log \frac{\text{Nombre de jours}}{\text{Nombre de jours où } S \text{ est présent}}$

Nous avons supposé que les sujets avec les plus grandes valeurs *Tf·Idf* correspondaient à des évènements. L'intuition derrière ce critère est de donner plus d'importance aux sujets qui ne sont pas fréquemment discutés au cours d'une période. Cette mesure pénalise les sujets fréquents apparaissant sur plusieurs jours.

6. Expérimentations

Dridi (2014) détaille l'ensemble des expériences menées pour évaluer les techniques présentées aux sections précédentes ainsi que la détermination des seuils à utiliser dans les algorithmes. Nous présentons maintenant les principaux enseignements que nous en avons tirés.

6.1. *Regroupement*

Les algorithmes de *Soundex* et de translittération (section 4.1) regroupent efficacement les hashtags (écrits en alphabets latin et arabe) de même prononciation sauf pour certains tags courts que l'algorithme regroupait trop agressivement. Nous avons utilisé des expressions régulières afin de normaliser les hashtags correspondant à des dates. Ces méthodes ont regroupé les 12 218 hashtags en 9 033 clusters, soit une réduction d'environ 26 %.

Comme expliqué en section 4.2, avec $CoDBscan$, nous avons réussi à regrouper des hashtags sémantiquement similaires même s'ils n'avaient pas une prononciation semblable, ce qui était impossible avec *Soundex*. Nous avons proposé deux variations de $CoDBscan$: la première s'appuie sur la mesure de *NPMI* pour déterminer le degré de similarité entre deux hashtags (ou codes *Soundex*) tandis que la deuxième utilise les hyperliens partagés par une paire de hashtags (ou codes *Soundex*) comme mesure de similarité. Les résultats obtenus ont montré que ces deux mesures (*NPMI* et les hyperliens partagés) sont de bons indices pour déterminer la similarité entre les hashtags.

DBscan n'était pas efficace dans notre cas parce qu'il construisait un énorme cluster. Pour cette raison, nous avons proposé la fonction de la cohésion qui vérifie la similarité entre un hashtag et un cluster avant de l'ajouter. Afin de regrouper un grand nombre de hashtags, nous avons effectué des améliorations avec *Soundex*. L'évaluation des résultats était coûteuse en temps, il nous fallait des heures ou même des jours pour calculer les précisions des différentes variantes de $CoDBscan$: $CoDBscan_{npmi}$ qui intègre les valeurs de *NPMI* et $CoDBscan_{npmiWithSndx}$ qui y ajoute les résultats du regroupement avec le *Soundex* ; nous avons de plus testé avec $CoDBscan_{hyper}$ et $CoDBscan_{hyperWithSndx}$ qui intègre l'information des hyperliens possiblement regroupés avec le *Soundex*.

Nos variantes de $CoDBscan$ utilisent trois paramètres : ***MinPts***, ϵ, λ. Pour les variantes $CoDBscan_{npmi}$ et $CoDBscan_{npmiWithSndx}$, il est recommandé d'utiliser une valeur $\epsilon \in$ [0,6, 0,8], car avec une telle valeur nous avons réussi à regrouper le plus grand nombre de hashtags avec une forte valeur de précision. Tandis que pour les variantes de $CoDBscan_{hyper}$ et $CoDBscan_{hyperWithSndx}$, il recommandé d'utiliser une valeur $\epsilon \in$ [0,4, 0,6] où la valeur de $epsilon$ correspond à la mesure hyper qui considère les hyperliens partagés entre deux hashtags.

Nos résultats ont été évalués en les comparant avec des annotations manuelles effectuées par des experts. Ces annotations pourraient être exploitées par d'autres travaux dans le futur sous forme de données de référence pour déterminer la similarité entre certains hashtags. Cependant, certains termes peuvent toujours être regroupés car ils n'ont pas de dépendance temporelle. D'autres sont regroupés seulement par période. Par exemple, une relation sémantique entre *université de manouba* et *drapeau tunisien* ne sera valide qu'au cours de la période analysée dans ce travail. En dehors de cette période, il y a peu de chances qu'il faille considérer ces termes comme similaires.

Tâches	#H	# groupes	#H regroupés	*Pr*
Normalisation *Soundex* et date	12 218	9 033	12 218	96 %
$CoDBscan_{npmi}(\epsilon = 0{,}7)$	9 973	928	5 011	94 %
$CoDBscan_{npmiWithSndx}$ ($\epsilon = 0{,}6$)	10 929	908	6 633	90 %
$CoDBscan_{hyper}$ ($\epsilon = 0{,}5$)	6 008	1 213	3 556	92 %
$CoDBscan_{hyperWithSndx}(\epsilon = 0{,}4)$	7 431	925	4 660	91 %

Tableau 3. *Résultats obtenus par les méthodes de regroupement de hashtags : #H indique le nombre de hashtags,* Pr *est la précision calculée pour la normalisation sur les 20 clusters plus importants.*

Le tableau 3 récapitule les principaux résultats obtenus. Pour la tâche de normalisation, le nombre de hashtags regroupés est le même que celui de nombre de hashtags considérés car chaque hashtag possède un code *Soundex*. Par contre pour les variantes de $CoDBscan$, il est possible qu'un hashtag n'appartienne à aucun cluster.

6.2. *Détection*

Nous avons testé trois méthodes afin de déterminer l'ensemble des évènements intéressants au cours d'une période :

– clusters de hashtags obtenus par une normalisation à base de prononciation (flèche du centre sur le rectangle ⑤ dans la figure 1) ;

– clusters de mots obtenus par LDA à base de techniques statistiques (flèche de droite sur le rectangle ⑤ dans la figure 1) ;

– clusters de tweets obtenus par un algorithme qui détermine lui-même le nombre de clusters de tweets similaires (flèche de gauche sur le rectangle ⑤ dans la figure 1).

Nous avons testé trois critères (voir section 5) : la fréquence quotidienne, la variation définie par l'écart-type de la fréquence quotidienne et une mesure inspirée du *Tf·Idf*.

Nous ne retenons que les sujets dont la valeur correspond à un pic identifié par la formule donnée en section 5. Par exemple, en appliquant cette méthode sur les 9 033 hashtags normalisés nous en avons retenu 123 selon le critère de fréquence, 88 selon la variation et 81 selon le *Tf·Idf*.

Dans ce type de problème, il est difficile de déterminer la pertinence des résultats obtenus, car il n'y a pas de données de référence. Pour évaluer l'adéquation de chaque critère et vérifier si les sujets détectés correspondaient à des évènements, nous avons demandé à dix experts tunisiens au courant des évènements en Tunisie de valider nos résultats. L'annotation des évènements a été faite par l'intermédiaire d'un site Web [5] que nous avons créé. Une fois connecté, l'expert a reçu une liste de sujets et les dates saillantes de chaque sujet et il a dû reconnaître le sujet référé par des hashtags et juger s'il s'agissait d'un évènement ou non.

Le tableau 4 résume les résultats des annotations effectuées sur les regroupements d'évènements :

– la normalisation des hashtags donne une bonne précision (nombre de vrais évènements détectés), mais un même sujet est souvent représenté par plusieurs clusters (tableau 3), ceci conduit à l'obtention de sujets similaires. Ce problème provient du fait que le sujet n'est représenté que par les termes (hashtags) avec une prononciation similaire. Le rappel n'est donc pas très bon, même s'il est difficile à évaluer précisément faute de référence. Un groupement typique est : `rvolution, revoliton2; revoltion, revolution, rrevolution; révolution, revolution2` ;

– les groupes de termes par LDA sont moins nombreux, mais un des problèmes reste le choix du nombre de sujets discutés. Le vocabulaire utilisé dans notre corpus est très riche, beaucoup de mots sont inventés par les utilisateurs. Malgré le stemming et la lemmatisation utilisés pour normaliser les termes, nous n'avons pas réussi à reconnaître certains mots, tels que des mots vides censés être supprimés. Un groupe typique est : `drapeau, faculté, salafistes, étudiant, manouba, lettre, ali, salafiste, police` ;

– le groupement de tweets par $CoDBscan$ combinés avec *NPMI* et les hyperliens nous semble être la méthode est la plus efficace pour déterminer l'ensemble des évènements intéressants. Presque tous les sujets identifiés sont distincts sans devoir fixer au préalable le nombre de sujets. Un exemple de ces groupes est : `#concours, #emploi, #interim, #jeune, #jobs, #programme, #rec, #recrutement, #chômage, #travail`.

On constate également que le critère inspiré du *Tf·Idf* est le plus efficace pour déterminer l'ensemble des évènements intéressants, et ce pour tous les types de regroupements.

5. `http://rali.iro.umontreal.ca:8080/dridihou/`

	Fréquence		Variation		*Tf·Idf*	
	# év	*Pr*	# év	*Pr*	# év	*Pr*
Hashtags normalisés	123	0,64	88	0,82	81	0,95
Groupes termes par *LDA*	53	0,73	67	0,87	74	0,92
Gr. tweets + $CoDBscan_{npmiWithSndx}$	98	0,61	52	0,80	46	0,93
Gr. tweets + $CoDBscan_{hyperWithSndx}$	104	0,54	51	0,69	46	0,94

Tableau 4. *Résultats des annotations effectués par 10* experts*. Pour chaque groupe et méthode d'identification des évènements, on indique le nombre d'évènements annotés (# év) avec la précision obtenue (*Pr*).*

7. Travaux futurs

À l'avenir, nous comptons tester ces méthodes sur d'autres types de données. Nous avons déjà commencé à extraire des données envoyées à partir du Québec. Cette fois-ci, nous avons profité de paramètres de géolocalisation pour extraire ces tweets. Comme mentionné précédemment, ces paramètres ne sont pas fonctionnels pour le cas de la Tunisie. Grâce à l'option de géolocalisation, nous n'avons pas eu à définir des mots-clés pour extraire les données et nous avons pu extraire un grand nombre de tweets, 4 millions entre février et juin 2014.

Sauf pour les méthodes utilisées dans la tâche de normalisation, notre système pourrait en principe fonctionner sur le corpus du Québec quoiqu'il faudra optimiser les temps de calcul qui seront prohibitifs sur des données aussi volumineuses. En travaillant sur ce corpus qui devrait contenir des tweets écrits en français et en anglais, il sera préférable d'utiliser une autre variante de *Soundex*, car celui que nous avons présenté ici est adapté aux textes écrits par des Tunisiens. Il sera toutefois inutile de translittérer les hashtags pour un corpus écrit entièrement en alphabet latin. En travaillant uniquement en anglais et français, nous pourrons profiter de ressources linguistiques (dictionnaires monolingues et bilingues, thesaurus, etc.) supportant ces langues.

Les méthodes proposées pour déterminer les termes similaires faisaient une distinction entre les mots et les hashtags trouvés dans les tweets. Aucune méthode ne servait à la création des clusters contenant à la fois des mots et des hashtags similaires. L'intuition derrière nos méthodes est de vérifier l'apport de chacun de ces éléments (mots et hashtags) pour déterminer les évènements au cours d'une période. Cependant, il est évident qu'il existe des hashtags et des mots sémantiquement similaires. Ainsi, nous pourrions proposer une méthode permettant de regrouper les hashtags et les mots.

Les résultats présentés dans ce travail prouvent que les hyperliens sont des éléments importants pour déterminer la similarité entre les hashtags. Rappelons que nous avons considéré que plus les hashtags partagent les mêmes hyperliens, plus ils sont similaires. Néanmoins, tout comme il y a des termes similaires, il y a aussi des hy-

perliens similaires (référant à un même sujet). Il serait donc intéressant de regrouper les hyperliens similaires ce qui permettait d'améliorer le regroupement de hashtags. De même, ils peuvent être utilisés comme indices pour identifier les évènements où chaque groupe d'hyperliens réfère à un sujet.

Dans cet article, nous n'avons pas distingué les tweets discutant d'un évènement de ceux qui donnent d'autres types d'informations. Comme extension, nous avons imaginé développer un classificateur permettant de déterminer si un tweet porte sur un évènement ou non. Pour ce faire, nous pourrions nous appuyer sur d'autres aspects que le seul contenu du tweet, par exemple tenir compte des utilisateurs mentionnés dans le tweet, est-ce une réponse à un autre tweet ? Est-ce un retweet ?

8. Conclusion

Dans ce travail, nous nous sommes intéressés à la détection des sujets ayant suscité l'intérêt des utilisateurs au cours d'une période donnée. Pour ce faire, nous avons eu recours aux données de médias sociaux qui constituent un excellent moyen pour les internautes pour partager leurs idées, donner leur avis et diffuser des nouvelles. Ces aspects conduisent à l'accumulation de données qui reflètent les préoccupations des utilisateurs en temps réel.

Notre première tâche dans le processus de détection d'évènements consiste à regrouper les termes similaires ou discutant du même sujet. Nous avons proposé des méthodes pour regrouper les hashtags avec une prononciation semblable. Pour ce faire, nous avons utilisé l'algorithme phonétique *Soundex*, que nous avons adapté au dialecte tunisien, pour attribuer le même code aux termes qui se prononcent de la même façon. Comme un même hashtag peut être écrit soit en alphabet latin soit en alphabet arabe, nous avons proposé un algorithme pour regrouper un hashtag écrit en alphabet arabe avec ses correspondants en alphabet latin. Pour réunir les hashtags référant à un même sujet, mais ne se prononçant pas de la même façon, nous avons modifié l'algorithme *DBscan* pour regrouper les hashtags indépendamment de la langue avec laquelle ils étaient écrits. Pour déterminer la similarité entre les hashtags, nous avons profité d'informations trouvées dans les tweets telles que la cooccurrence et les hyperliens partagés entre les hashtags.

Nous avons utilisé l'algorithme LDA pour regrouper les mots qui portent sur le même sujet. Nous avons regroupé dans un seul document les tweets partageant les mêmes hashtags afin d'améliorer les résultats. L'absence de données de référence nous a obligés à évaluer manuellement les résultats de nos méthodes de regroupement de hashtags. L'évaluation, effectuée par des experts familiarisés avec les évènements qui se sont déroulés en Tunisie, a révélé une précision qui dépassait souvent 90 %.

Par la suite, nous avons utilisé ces clusters pour déterminer les sujets saillants ou les évènements. Pour ce faire, nous avons proposé deux méthodes. À partir des clusters obtenus par la normalisation des hashtags et de LDA, nous avons considéré que chaque cluster correspondait à un sujet. Un tweet porte sur un sujet si au moins l'un de ses

termes est présent. Cette tâche a permis d'obtenir la fréquence quotidienne de chaque sujet. Nous avons également expérimenté en regroupant les tweets similaires avec un algorithme de regroupement incrémental déterminant le nombre de clusters.

Comme tous les sujets ne sont pas nécessairement des évènements, nous avons adapté la méthode de Palshikar (2009) pour sélectionner les sujets saillants en évaluant trois critères : fréquence, écart-type, *Tf·Idf*. Notre évaluation a montré que le critère *Tf·Idf* est le plus adéquat pour cette tâche.

9. Bibliographie

Allan J., *Topic detection and tracking : event-based information organization*, vol. 12, Kluwer Academic Publishers, 2002.

Allan J., Carbonell J., Doddington G., Yamron J., Yang Y., « Topic detection and tracking pilot study final report », 1998.

Becker H., Naaman M., Gravano L., « Beyond trending topics : Real-world event identification on Twitter », *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM11)*, 2011.

Blei D., Ng A., Jordan M., « Latent Dirichlet Allocation », *Journal of Machine Learning Research*, vol. 3, p. 993-1022, 2003.

Boujelbane R., « Génération des corpus en dialecte tunisien pour la modélisation de langage d'un système de reconnaissance », *Actes de la 15e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'2013)*, Les Sables d'Olonne, France, p. 206-216, 2013.

Bouma G., « Normalized (pointwise) mutual information in collocation extraction », *Proceedings of the Biennial GSCL Conference*, p. 31-40, 2009.

Chakrabarti D., Punera K., « Event summarization using tweets », *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, p. 66-73, 2011.

Church K., Hanks P., « Word association norms, mutual information, and lexicography », *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics, p. 76-83, 1989.

Clark E., Araki K., « Text normalization in social media : progress, problems and applications for a pre-processing system of casual English », *Procedia-Social and Behavioral Sciences*, vol. 27, p. 2-11, 2011.

Dridi H. e., Détection d'évènements à partir de Twitter, PhD thesis, Université de Montréal, Montréal, Canada, oct, 2014.

Ester M., Kriegel H.-P., Sander J., Xu X., « A density-based algorithm for discovering clusters in large spatial databases with noise. », *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, p. 226-231, 1996.

Farzindar A., Khreich W., « A survey of techniques for event detection in Twitter », *Computational Intelligence (Early View)*p. 33 pages, sept, 2013.

Jianshu W., Bu-Sung L., « Event Detection in Twitter », *in* L. Adamic, R. Baeza-Yates, S. Counts (eds), *ICWSM*, The AAAI Press, p. 401-408, 2011.

Kaiser G., *A friendly guide to wavelets*, Springer, 2011.

Kleinberg J., « Bursty and hierarchical structure in streams », *Data Mining and Knowledge Discovery*, vol. 7, n° 4, p. 373-397, 2003.

Kontostathis A., Galitsky L., Pottenger W., Roy S., Phelps D., « A survey of emerging trend detection in textual data mining », *Survey of Text Mining*, Springer, p. 185-224, 2004.

Kwak H., Lee C., Park H., Moon S., « What is Twitter, a social network or a news media ? », *Proceedings of the 19th international conference on World wide web*, ACM, p. 591-600, 2010.

Makkonen J., Ahonen-Myka H., Salmenkivi M., « Simple semantics in topic detection and tracking », *Information Retrieval*, vol. 7, n° 3, p. 347-368, 2004.

McCallum A., « Mallet : A machine learning for language toolkit », 2002, http ://mallet.cs.umass.edu.

Ozdikis O., Senkul P., Oguztuzun H., « Semantic Expansion of Tweet Contents for Enhanced Event Detection in Twitter », *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, IEEE, p. 20-24, 2012a.

Ozdikis O., Senkul P., Oguztuzun H., « Semantic Expansion of Tweet Contents for Enhanced Event Detection in Twitter », *ASONAM*, IEEE Computer Society, p. 20-24, 2012b.

Palshikar G., Simple algorithms for peak detection in time-series, Technical report, TRDDC, 2009.

Petrovic S., Osborne M., Lavrenko V., « Streaming First Story Detection with application to Twitter », *HLT-NAACL*, The Association for Computational Linguistics, p. 181-189, 2010.

Pustejovsky J., Lee K., Bunt H., Romary L., « ISO-TimeML : An International Standard for Semantic Annotation », *in* E. L. R. A. (ELRA) (ed.), *Proceedings of the 7th International Conference on Language Re- sources and Evaluation (LREC'10)*, Valletta, Malta, May, 2010.

Russell R., « Soundex coding system », *United States Patent*, 1918.

Sakaki T., Okazaki M., Matsuo Y., « Earthquake shakes Twitter users : real-time event detection by social sensors », *Proceedings of the 19th international conference on World wide web*, ACM, p. 851-860, 2010.

Shamma D. A., Kennedy L., Churchill E., « Statler : Summarizing media through short-message services », *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW10)*, p. 551-552, 2010.

Sproat R., Black A., Chen S., Kumar S., Ostendorf M., Richards C., « Normalization of non-standard words », *Computer Speech & Language*, vol. 15, n° 3, p. 287-333, 2001.

Sutton J., Palen L., Shklovski I., « Backchannels on the front lines : Emergent uses of social media in the 2007 southern California wildfires », *Proceedings of the 5th International ISCRAM Conference*, Washington, DC, p. 624-632, 2008.

Zhang K., Zi J., Wu L., « New event detection based on indexing-tree and named entity », *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, p. 215-222, 2007.

Code-Mixing in Social Media Text

The Last Language Identification Frontier?

Amitava Das* — **Björn Gambäck****

* *NITT University, Neemrana, Rajasthan 301705, India*
amitava.santu@gmail.com

** *Norwegian University of Science and Technology, 7491 Trondheim, Norway*
gamback@idi.ntnu.no

ABSTRACT. *Automatic understanding of noisy social media text is one of the prime present-day research areas. Most research has so far concentrated on English texts; however, more than half of the users are writing in other languages, making language identification a pre-requisite for comprehensive processing of social media text. Though language identification has been considered an almost solved problem in other applications, language detectors fail in the social media context due to phenomena such as code-mixing, code-switching, lexical borrowings, Anglicisms, and phonetic typing. This paper reports an initial study to understand the characteristics of code-mixing in the social media context and presents a system developed to automatically detect language boundaries in code-mixed social media text, here exemplified by Facebook messages in mixed English-Bengali and English-Hindi.*

RÉSUMÉ. *La compréhension automatique du texte bruyant des médias sociaux est l'un des secteurs de recherche contemporaine principaux. Jusqu'ici, la plupart des recherches se sont concentrées sur les textes en anglais ; mais plus de la moitié des utilisateurs écrivent dans d'autres langues, ce qui rend l'identification de la langue préalable au traitement complet du texte des médias sociaux. Bien que l'identification de la langue ait été considérée comme un problème presque résolu dans d'autres applications, les détecteurs de langue échouent dans le contexte des médias sociaux, et cela est dû aux phénomènes tels que le mélange et l'alternance de code linguistique, les emprunts lexicaux, les anglicismes et la dactylographie phonétique. Cet article présente une étude initiale pour comprendre les caractéristiques de mélange des codes dans le contexte des médias sociaux ainsi qu' un système développé pour détecter automatiquement les barrières linguistiques en texte «code-mélangé» de médias sociaux, ici illustrées par des messages de Facebook en mixte anglais-bengali et anglais-hindi.*

KEYWORDS: *Code-mixing, code-switching, social media text, language identification.*

MOTS-CLÉS : *Mélange et alternance de code linguistique, textes des médias sociaux, identification de la langue.*

TAL Volume 54 – n° 3/2013, pages 41 à 64

1. Introduction

The evolution of social media texts, such as Twitter and Facebook messages, has created many new opportunities for information access and language technology, but also many new challenges, in particular since this type of text is characterized by having a high percentage of spelling errors and containing creative spellings (*"gr8"* for *'great'*), phonetic typing, word play (*"goooood"* for *'good'*), abbreviations (*"OMG"* for *'Oh my God!'*), Meta tags (*URLs, Hashtags*), and so on. So far, most of the research on social media texts has concentrated on English, whereas most of these texts now are in non-English languages (Schroeder, 2010). Another study (Fischer, 2011) provides an interesting insight on Twitter language usages from different geospatial locations. It is clear that even though English still is the principal language for web communication, there is a growing need to develop technologies for other languages. However, an essential prerequisite for any kind of automatic text processing is to first identify the language in which a specific text segment is written. The work presented here will in particular look at the problem of word-level identification of the different languages used in social media texts. Available language detectors fail for social media text due to the style of writing, despite a common belief that language identification is an almost solved problem (McNamee, 2005).

In social media, non-English speakers do not always use Unicode to write in their own language, they use phonetic typing, frequently insert English elements (through code-mixing and Anglicisms), and often mix multiple languages to express their thoughts, making automatic language detection in social media texts a very challenging task. All these language mixing phenomena have been discussed and defined by several linguists, with some making clear distinctions between phenomena based on certain criteria, while others use 'code-mixing' or 'code-switching' as umbrella terms to include any type of language mixing (Auer, 1999; Muysken, 2000; Gafaranga and Torras, 2002; Bullock *et al.*, 2014), as it is not always clear where borrowings/Anglicisms stop and code-mixing begins (Alex, 2008). In the present paper, 'code-mixing' will be the term mainly used (even though 'code-switching' thus is equally common). Specifically, we will take 'code-mixing' as referring to the cases where the language changes occur inside a sentence (which also sometimes is called intra-sentential code-switching), while we will refer to 'code-switching' as the more general term and in particular use it for inter-sentential phenomena.

Code-mixing is much more prominent in social media than in more formal texts, as in the following examples of mixing between English and Bengali (the language spoken in Eastern India and Bangladesh), where the Bengali segments (bold) are written using phonetic typing and not Unicode. Each example fragment (in italics) is followed by the corresponding English gloss on the line after it.

[1] ***ki korle ekta darun*** *hot gf* ***pao jabe setai bujte parchina***
What do I need to do to have a hot girlfriend, I'm unable to figure that out,
please help seniors.
please help seniors.

[2] ***Ami*** *hs* ***a 65% paya6i n madhyamik a 88%***
I got 65% in HS and 88% in Madhyamik
..but ju ***te*** *physics* ***nya porte chai***
..but I wanted to study physics at JU
..but am nt eligbl 4 dat course bcoz of mah 12th no
..but I am not eligible for that course because of my 12th mark
..but ***amr*** *wbjee* ***te*** *rank 88 ..ju* ***te sb kichu pa66i***
..but my WBJEE rank is 88 ..I am taking engineering at JU
..but ***ami*** *engineering* ***porte chai na*** *..i love physics and*
..but I don't want to study engineering ..I love physics and
ju ***r mto kno*** *clg* ***thaka porte chai.*** ***kao ki*** *hlp* ***krbe***
wanted to study at a college like JU. Can anybody help me
..wbjee rank ***dakhia ki*** *ju* ***te*** *physics* ***paoa jbe?*** *plz hlp.*
..Can I get entrance waiver with my WBJEE rank? Please help.

In Example 2, "HS" stands for 'higher secondary 10' and "JU" is Jadavpur University, while "Madhyamik" is the 10th grade exam in the Eastern Indian state of West Bengal, "12 no" refers to the 12th grade math mark, and "WBJEE" means the 'West Bengal Joint Entrance Examination' (the exam for admissions to engineering courses).

The remainder of the paper is laid out as follows: the next section discusses the concept of code-switching and some previous studies on code-mixing in social media text. Then Section 3 introduces the data sets that have been used in the present work for investigating code-mixing between English and Hindi as well as between English and Bengali. The data stem from two different Indian universities' campus-related billboard postings on Facebook. Section 4 describes the various methods used for word-level language detection, based respectively on character n-grams, dictionaries, and support vector machines. The actual experiments on language detection are reported in Section 5. Finally, Section 6 sums up the discussion and points to some areas of future research.

2. Background and Related Work

In the 1940s and 1950s, code-switching was often considered a sub-standard use of language. However, since the 1980s it has generally been recognized as a natural part of bilingual and multilingual language use. Linguistic efforts in the field have mainly concentrated on the sociological and conversational necessity behind code-switching and its linguistic nature (Muysken, 1995; Auer, 1984), dividing it into various sub-categories such as *inter-* vs *intra-sentential switching* (depending on whether it occurs outside or inside sentence or clause boundaries); *intra-word* vs *tag switching* (if the switching occurs within a word, for example at a morpheme boundary, or by inserting a tag phrase or word from one language into another), and on whether the switching is an act of identity in a group or if it is competence-related (that is, a consequence of a lack of competence in one of the languages).

Following are some authentic examples of each type of code-switching from our English-Bengali corpus (the corpus is further described in Section 3). Again, Bengali segments are in boldface and each example fragment is followed by its corresponding English gloss on a new line. In the intra-word case (Example 6), the plural suffix of *admirer* has been Bengalified to *der*.

[3] Inter-sentential:
*Fear cuts deeper than sword .... **bukta fete jachche** ... :(*
Fear cuts deeper than a sword .... it seems my heart will blow up ... :(

[4] Intra-sentential:
***dakho sune 2mar kharap lagte pare** but it is true that u r confused.*
You might feel bad hearing this but it is true that you are confused.

[5] Tag:
ami majhe majhe** fb **te on hole ei** confession page **tite aasi.
While I get on facebook I do visit the confession page very often.

[6] Intra-word:
tomar osonkkhho** admirer **der** **modhhe ami ekjon nogonno manush
Among your numerous admirer-s I am the negligible one

2.1. *Characteristics of Code-Mixing*

The first work on processing code-switched text was carried out over thirty years ago by Joshi (1982), while efforts in developing tools for automatic language identification started even earlier (Gold, 1967). Still, the problem of applying those language identification programs to multilingual code-mixed texts has only started to be addressed in very recent time. However, before turning to that topic, we will first briefly discuss previous studies on the general characteristics of code-mixing in social media text, and in particular those on the reasons for users to mix codes, on the types and the frequencies of code-mixing, and on gender differences.

Clearly, there are (almost) as many reasons for why people code-switch as there are people code-switching. However, several studies of code-switching in different type of social media texts indicate that social reasons might be the most important, with the switching primarily being triggered by a need in the author to mark some in-group membership. So did Sotillo (2012) investigate the types of code-mixing occurring in short text messages, analysing an 880 SMS corpus, indicating that the mixing often takes place at the beginning of the messages or through simple insertions, and mainly to mark in-group membership — which also Bock (2013) points to as the main reason for code-mixing in a study on chat messages in English, Afrikaans and isiXhosa. Similar results were obtained by Xochitiotzi Zarate (2010) in a study on English-Spanish SMS text discourse (although based on only 42 text messages), by Shafie and Nayan (2013) in a study on Facebook comments (in Bahasa Malaysia and English), and by Negrón Goldbarg (2009) in a small study of code-switching in the emails of five Spanish-English bilinguals. However, this contrasts with studies on

Chinese-English code-mixing in Hong Kong by Li (2000) and in Macao by San (2009) with both indicating that code-switching in those highly bilingual societies mainly is triggered by linguistic motivations, with social motivations being less salient.

Two other topics that have been investigated relate to the frequency and types of code-switching in social media. Thus Dewaele (2008; 2010) claimed that "strong emotional arousal" increases the frequency of code-switching. Johar (2011) investigated this, showing that an increased amount of positive smileys indeed was used when code-switching. On the types of switching, San's (2009) study, which compared the switching in blog posts to that in the spoken language in Macao, reported a predominance of inter-sentential code-switching. Similarly, Hidayat (2012) noted that facebookers tend to mainly use inter-sentential switching (59%) over intra-sentential (33%) and tag switching (8%), and reports that 45% of the switching was instigated by real lexical needs, 40% was used for talking about a particular topic, and 5% for content clarification. In contrast, our experience of code-switching in Facebook messages is that intra-sentential switching tends to account for more than half of the cases, with inter-sentential switching only accounting for about $1/3$ of the code-switching (Das and Gambäck, 2014).

Furthermore, a few studies have looked at differences in code-switching behaviour between groups and types of users, in particular investigating gender-based ones. Kishi Adelia (2012) manually analysed the types and functions of code-switching used by male and female tweeters, but on a very small dataset: only 100 tweets from 20 participants. The results indicate that male Indonesian students predominantly prefer intra-sentential code-switching and use it to show group membership and solidarity, while female students rather tend to utilize inter-sentential code-switching in order to express feelings and to show gratitude. Ali and Mahmood Aslam (2012) also investigated gender differences in code-switching, in a small SMS corpus, indicating that Pakistani female students have a stronger tendency than males to mix English words into their (Urdu) texts.

2.2. *Automatic Analysis of Code-Switching*

Turning to the work on automatic analysis of code-switching, there have been some related studies on code-mixing in speech (e.g., Chan *et al.*, 2009; Solorio *et al.*, 2011; Weiner *et al.*, 2012). Solorio and Liu (2008a) tried to predict the points inside a set of spoken Spanish-English sentences where the switch between the two languages occur, while (Rodrigues and Kübler, 2013) looked at part-of-speech tagging for this type of data, as did (Solorio and Liu, 2008b), in part by utilising a language identifier as a pre-processing step, but with no significant improvement in tagging accuracy. Notably, these efforts have mainly been on artificially generated speech data, with the simplification of only having 1–2 code-switching points per utterance. The spoken Spanish-English corpus used by Solorio and Liu (2008b) is a small exception, with 129 intra-sentential language switches.

Previous work on text has mainly been on identifying the (one, single) language (from several possible languages) of documents or the proportion of a text written in a language, often restricted to 1–2 known languages; so even when evidence is collected at word-level, evaluation is at document-level (Prager, 1997; Singh and Gorla, 2007; Yamaguchi and Tanaka-Ishii, 2012; Rodrigues, 2012; King and Abney, 2013; Lui *et al.*, 2014). Other studies have looked at code-mixing in different types of short texts, such as information retrieval queries (Gottron and Lipka, 2010) and SMS messages (Rosner and Farrugia, 2007), or aimed to utilize code-mixed corpora to learn topic models (Peng *et al.*, 2014) or user profiles (Khapra *et al.*, 2013).

Most closely related to the present work are the efforts by Carter (2012), by Nguyen and Doğruöz (2013), by Lignos and Marcus (2013), and by Voss *et al.* (2014). Nguyen and Doğruöz investigated language identification at the word-level on randomly sampled mixed Turkish-Dutch posts from an online forum, mainly annotated by a single annotator, but with 100 random posts annotated by a second annotator. They compared dictionary-based methods to language models, and with adding logistic regression and linear-chain Conditional Random Fields (CRF). The best system created by Nguyen and Doğruöz (2013) reached a high word-level accuracy (97.6%), but with a substantially lower accuracy on post-level (89.5%), even though 83% of the posts actually were monolingual.

Similarly, Lignos and Marcus (2013) also only addressed the bi-lingual case, looking at Spanish-English Twitter messages (tweets). The strategy chosen by Lignos and Marcus is interesting in its simplicity: they only use the ratio of the word probability as information source and still obtain good results, the best being 96.9% accuracy at the word-level. However, their corpora are almost monolingual, so that result was obtained with a baseline as high as 92.3%.

Voss *et al.* (2014) on the other hand worked on quite code-mixed tweets (20.2% of their test and development sets consisted of tweets in more than one language). They aimed to separate Romanized Moroccan Arabic (Darija), English and French tweets using a Maximum Entropy classifier, achieving F-scores of .928 and .892 for English and French, but only .846 for Darija due to low precision.

Carter collected tweets in five different languages (Dutch, English, French, German, and Spanish), and manually inspected the multilingual micro-blogs for determining which language was the dominant one in a specific tweet. He then performed language identification at post-level only, and experimented with a range of different models and a character n-gram distance metric, reporting a best overall classification accuracy of 92.4% (Carter, 2012; Carter *et al.*, 2013). Evaluation at post-level is reasonable for tweets, as Lui and Baldwin (2014) note that users who mix languages in their writing still tend to avoid code-switching within a tweet. However, this is not the case for the chat messages that we address in the present paper.

Code-switching in tweets was also the topic of the shared task at the recent First Workshop on Computational Approaches to Code Switching for which four different code-switched corpora were collected from Twitter (Solorio *et al.*, 2014). Three

of these corpora contain English-mixed data from Nepalese, Spanish and Mandarin Chinese, while the fourth corpus consists of tweets code-switched between Modern Standard Arabic and Egyptian Arabic. Of those, the Mandarin Chinese and (in particular) the Nepalese corpora exhibit very high mixing frequencies. This could be a result of the way the corpora were collected: the data collection was specifically targeted at finding code-switched tweets (rather than finding a representative sample of tweets). This approach to the data collection clearly makes sense in the context of a shared task challenge, although it might not reflect the actual level of difficulty facing a system trying to separate "live" data for the same language pair.

3. The Nature of Code-Switching in Social Media Text

According to the Twitter language map, Europe and South-East Asia are the most language-diverse areas of the ones currently exhibiting high Twitter usage. It is likely that code-mixing is frequent in those regions, where languages change over a very short geospatial distance and people generally have basic knowledge of the neighbouring languages. Here we will concentrate on India, a nation with close to 500 spoken languages (or over 1600, depending on what is counted as a language and what is treated as a dialect) and with some 30 languages having more than 1 million speakers. India has no national language, but 22 languages carry official status in at least parts of the country, while English and Hindi are used for nation-wide communication. Language diversity and dialect changes instigate frequent code-mixing in India, and already in 1956 the country's Central Advisory Board on Education adopted what is called the "three-language formula", stating that three languages shall be taught in all parts of India from the middle school and upwards (Meganathan, 2011). Hence, Indians are multi-lingual by adaptation and necessity, and frequently change and mix languages in social media contexts. Most frequently, this entails mixing between English and Indian languages, while mixing Indian languages is not as common, except for that Hindi as the primary nation-wide language has high presence and influence on the other languages of the country.

English-Hindi and English-Bengali language mixing were selected for the present study. These language combinations were chosen as Hindi and Bengali are the two largest languages in India in terms of first-language speakers (and 4th and 7th worldwide, respectively). To understand the relation between topic and code-mixing, we collected data including both formal and informal topics. The formal data mainly come from placement forums, where people discuss and exchange information about various companies, selection processes, interview questions, and so on. The informal data is generally on fun topics such as on-campus love confession, on-campus matrimonial, etc. For the English-Bengali pair, the data came from Jadavpur University, which is located in Eastern India where the native language of most of the students is Bengali. For English-Hindi, the data came from the Indian Institute of Technology Bombay (IITB), an institution located in the West of India where Hindi is the most common language.

Language Pair	Facebook Group	Messages	Type
English	JU Confession	5,040	Informal
—	JU Matrimonial	4,656	Informal
Bengali	Placement 2,013 Batch	500	Formal
English	IITB Confession	1,676	Informal
—	IITB Compliments	1,717	Informal
Hindi	Tech@IITB	631	Formal

Table 1. *Details of corpus collection.*

Number of	English–Bengali	English–Hindi
Sentences	24,216	8,901
Words	193,367	67,402
Unique Tokens	100,227	40,240

Table 2. *Corpus size statistics.*

3.1. *Data Acquisition*

Various campus Facebook groups were used for the data acquisition, as detailed in Table 1. The data was annotated by five annotators, using GATE (Bontcheva *et al.*, 2013), as annotation tool. The two corpora (English-Bengali and English-Hindi) were then each split up into training (60%), development (20%), and test (20%) sets. Table 2 presents corpus statistics for both language pairs.

None of the annotators was a linguist. Out of the five, three were native Bengali speakers who knew Hindi as well. The other two annotators were native Hindi speakers not knowing Bengali. Hence all the English-Hindi data was annotated by all the five annotators, while the English-Bengali data was annotated only by the three native speakers. Among the annotators, four (both the Hindi speakers and two of the Bengali speakers) were college students and the fifth a Bengali speaking software professional.

The annotators were instructed to tag language at the word-level with the tagset displayed in Table 3. Each tag was accompanied by some examples. The `univ` tag stands for emoticons `(:)`, `:(`, etc.) and characters such as `"`, `’`, `>`, `!`, and `@`, while `undef` is for the rest of the tokens and for hard to categorize or bizarre things. The overall annotation process was not very ambiguous and annotation instruction was also straight-forward. The inter-annotation agreement was above 98% and 96% (average for all the tags) for English-Bengali and English-Hindi, respectively, with *kappa* measures of 0.86 and 0.82 for the two language pairs.

Tag	Description	Examples
en	English word	dear, help, please
en+bn_suffix	English word + Bengali suffix ("*Engali*")	world-er (*of this world*)
en+hi_suffix	English word + Hindi suffix ("*Engdi*")	desh-se (*from country*)
bn	Bengali word	lokjon (*people*), khub (*very*)
bn+en_suffix	Bengali word + English suffix ("*Benglish*")	addaing (*gossiping*)
hi	Hindi word	pyar (*love*), jyada (*more*)
hi+en_suffix	Hindi word + English suffix ("*Hinglish*")	jugading (*making arrangements*)
ne	Named Entity (NE)	Kolkata, Mumbai
ne+en_suffix	NE + English suffix	Valentine's, Ram's
ne+bn_suffix	NE + Bengali suffix	rickshaw-r (*of rickshaw*), mahalayar (*about mahalaya*)
ne+hi_suffix	NE + Hindi suffix	Tendulkarka (*Tendulkar's*), Riaki (*Ria's*)
acro	Acronyms	JU (*Jadavpur University*), UPA
acro+en_suffix	Acronym + English suffix	VC's, IITs
acro+bn_suffix	Acronym + Bengali suffix	JUr (*of JU*)
acro+hi_suffix	Acronym + Hindi suffix	IITka (*of IIT*)
univ	Universal	", ', >, !, @, ..., :), :(
undef	Undefined	rest of the tokens, hard to categorize or strange things

Table 3. *Word-level code-mixing annotation tagset.*

Some ambiguous cases are "*Bengali word + English suffix*" and "*Hindi word + English suffix*", that is, cases of *Benglish* and *Hinglish*. Other problems were related to determining where code-mixing ends and borrowing (Anglicism) begins, as exemplified by the English word "*glass*" (as in drinking glass: a container made of glass for holding liquids while drinking). The concept of "*glass*" was borrowed during the British colonisation in India. Though there are symbolic Indian words that have been synthesized later on to cover the same concept, Indian dictionaries still consider the original word-form "*glass*" (transliterated into Indian languages) as a valid Indian word. However, the annotators sometimes labelled it as a foreign word, and hence an Anglicism.

Language Pair	Topic Type	Code-Switching Types			Total
		Intra	Inter	Word	
ENG-HND	**Informal**	54.95%	36.85%	8.2%	32.37%
	Formal	53.42%	39.88%	6.7%	8.25%
ENG-BNG	**Informal**	60.21%	32.09%	7.7%	58.82%
	Formal	60.61%	34.19%	5.2%	12.58%

Table 4. *Topic-wise code-switching and categorisation.*

3.2. *Types of Code-Switching*

The distribution of topic and code-switching is reported in Table 4, under the hypothesis that the base language is English with the non-English words (i.e., Hindi/Bengali) having been mixed in. Named entities and acronyms were treated as language independent, but assigned the language for multilingual categories based on suffixes. From the statistics, it is clear that people are much more inclined to use code-mixing or their own languages when writing on informal rather than more formal topics, where the mixing is only about 1/4 as frequent.

The 'total' percentage in Table 4 was calculated at the word level (so not on the number of sentences, but rather on the number of words in those sentences), that is, as in Equation 7.

$$\frac{\textit{total number of words found in non-English}}{\textit{total number of words in the corpus}} \quad [7]$$

The inter- and intra-sentential code-switching figures for each language-topic corpus were calculated automatically and based on the total code-switching found in the corpus: if the language of a sentence was fully tagged either as Bengali or Hindi, then that sentence was considered as a type of inter-sentential code-switching, and all words in that sentence contribute to the inter-sentential code-switching percentage. For word-internal code-mixing identification, only the "`* + * suffix`" tags were considered. Tag-mixing was not considered or annotated as it either is a semantic category or can be further described as a subtype of intra-sentential code-switching.

Suppose that the total number of non-English words in the ENG-BNG informal corpus is n. If the words present for each switching-type (that is, word-level, intra- and inter-sentential switching) are m_w, m_s and $(n - m_w - m_s)$, respectively, then the percentage of each switching category is calculated at word-level by Equation 8, intra-sentential code-switching by Equation 9, and inter-sentential by Equation 10.

$$\frac{m_w}{n} \quad [8]$$

$$\frac{m_s}{n} \quad [9]$$

$$\frac{(n - m_w - m_s)}{n} \quad [10]$$

For example, the total code-switching percentage of ENG–HND informal topic is 32.37%, which is the fraction of non-English words in that corpus.

A typical inter-sentential code-switching example from our 'informal' English-Bengali corpus is shown below.

[11]	*Yaar tu to,* GOD *hain.*	***tui JU te ki korchis?***	Hail	u	man!
	Dude you are GOD.	What you are doing in JU?	Hail	you	man!

This comment was written in three languages: English, Hindi (italics), and Bengali (boldface italics; "JU" is an abbreviation for Jadavpur University, but we hypothesized that named entities are language independent). The excerpt stems from the "JU Confession" corpus, which in general is an ENG-BNG group; however, it has a presence of 3–4% Hindi words mixed (due to Hindi being India's primary nation-wide language, as noted above). It is clear from the example how closely languages coexist in social media text, making language detection for this type of text a very complex task.

4. Word-Level Language Detection

The task of detecting the language of a text segment in mixed-lingual text remains beyond the capabilities of existing automatic language identification techniques (e.g., Beesley, 1988; Dunning, 1994; Cavnar and Trenkle, 1994; Damashek, 1995; Ahmed *et al.*, 2004). We tested some of the state-of-the-art language identification systems on our corpora and found that they in general fail to separate language-specific segments from code-switched texts. [1] Instead we designed a system based on well-studied techniques, namely character n-gram distance measures, dictionary-based information, and classification with support vector machines (SVM), as described in the present section. The actual experiments and results with this system are reported in Section 5, which also discusses ways to improve the system by adding post-processing.

1. The language identification systems tested were:
   - WiseGuys' `LibTextCat`: `software.wise-guys.nl/libtextcat`
   - Jelsma's `LanguageIdentifier`: `wiki.apache.org/nutch/LanguageIdentifier`
   - Shuyo's `LanguageDetectionLib`: `code.google.com/p/language-detection`
   - Xerox' `LanguageIdentifier`: `open.xerox.com/Services/LanguageIdentifier`
   - Lui's `langid.py`: `github.com/saffsd/langid.py`

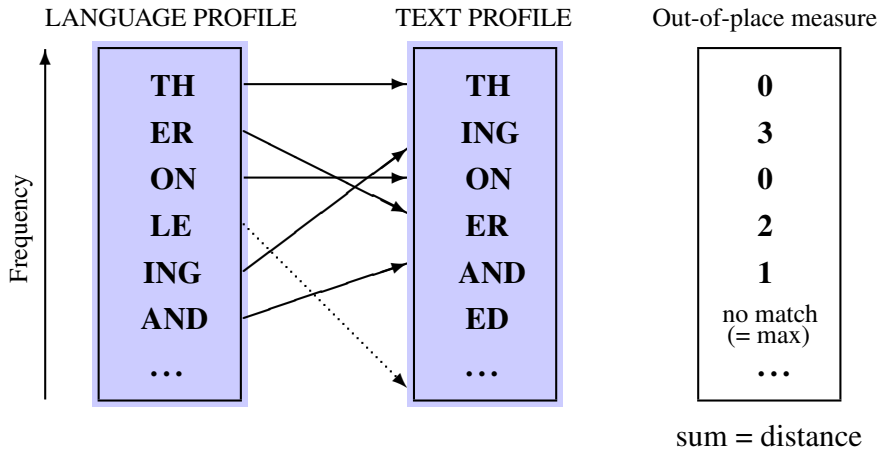

Figure 1. *Language detection by character n-gram frequency. Reproduced from Cavnar and Trenkle (1994).*

4.1. *N-Gram Language Profiling and Pruning*

The probably most well-known language detection system is *TextCat* (Cavnar and Trenkle, 1994; van Noord, 1997) which utilizes character-based n-gram models. The method generates language specific n-gram profiles from the training corpus sorted by their frequency. A similar text profile is created from the text to be classified, and a cumulative "out-of-place" measure between the text profile and each language profile is calculated, as illustrated in Figure 1. The measure determines how far an n-gram in one profile is from its place in the other profile. Based on that distance value, a threshold is calculated automatically to decide the language of a given text. This approach has been widely used and is well established in language identification (e.g., Beesley, 1988; Dunning, 1994; Teahan, 2000; Ahmed, 2005). Andersen (2012) also investigated n-gram based models, both in isolation and in combination with the dictionary-based detection described in the next section, as well as with a rule-based method utilising manually constructed regular expressions.

An n-gram model was adopted for the present task, too, but with a pruning technique to exclude uninformative n-grams during profile building. Common (high-frequency) n-grams for both language pairs are removed, as they are ambiguous and less discriminative. So is, for example, the bigram '**TO**' very common in all the three languages (English, Hindi, and Bengali), so less discriminative and has been excluded.

To achieve this, a weight ϕ_i^a is calculated for each n-gram γ_i in language l_a by the formula in Equation 12

$$\phi_i^a = \frac{f_i^a}{m_a} \quad [12]$$

where f_i^a is the frequency of the n-gram γ_i in language l_a and m_a the total number of n-grams in language l_a.

A particular n-gram γ_i is excluded if its discriminative power when comparing languages l_a and l_b is lower than an experimentally chosen threshold value θ, that is, if the condition in Equation 13 is true.

$$|\phi_i^a - \phi_i^b| \leq \theta \quad [13]$$

There are various trade-offs to consider when choosing between character n-grams and word n-grams, as well as when deciding on the values of n and θ, that is, the size of the n-grams and the discrimination threshold. Using Romanization for the Hindi and Bengali, and converting all text to lower-case, the alphabet of English is limited to 26 characters, so the set of possible character n-grams remains manageable for small values of n. The white-spaces between the words were kept for the n-gram creation, in order to distinctly mark word boundaries, but multiple white-spaces were removed.

We carried out experiments on the training data for $n = \{1, 2, 3, 4, 5, 6, 7\}$, and found 3-grams and 4-grams to be the optimum choices after performance testing through 10-fold cross validation, with $\theta = 0.2$. The value of θ was not varied: n-grams with the same presence in multiple languages are less discriminating. The presence ratio should be $> 2\%$, so that value was selected for θ. N-gram pruning helps reduce the time it takes the system to converge by a factor 5 and also marginally increases performance (by 0.5).

4.2. *Dictionary-Based Detection*

Use of most-frequent-word dictionaries is another established method in language identification (Alex, 2008; Řehůřek and Kolkus, 2009). We incorporated a dictionary-based language detection technique for the present task, but were faced with a few challenges for the dictionary preparation, in particular since social media text is full of noise. A fully edited electronic dictionary may not have all such distorted word forms as are used in these texts (e.g., '*gr8*' rather than '*great*'). Therefore a lexical normalisation dictionary (Han *et al.*, 2012; Baldwin, 2012) prepared for Twitter was used for English.

Unfortunately, no such dictionary is available for Hindi or Bengali, so we used the Samsad English-Bengali dictionary (Biśvās, 2000; Digital South Asia Library, 2006). The Bengali part of the Samsad dictionary is written in Unicode, but in our corpus

the Bengali texts are written in transliterated/phonetic (Romanized) form. Therefore the Bengali lexicon was transliterated into Romanized text using the Modified-Joint-Source-Channel model as described by Das *et al.* (2010). The same approach was taken for the Hindi dictionary creation, using Hindi WordNet (Narayan *et al.*, 2002; Center for Indian Language Technology, 2013).

In order to capture all the distorted word forms for Hindi and Bengali, an edit distance (Levenshtein, 1966) method was adopted. A Minimum Edit Distance (MED) of ± 3 was used as a threshold (chosen experimentally). The general trend in dictionary-based methods is to keep only high-frequency words, but that is for longer texts, and surely not for code-mixing situations. Our language detection solution is targeted at the word-level and for short texts, so we cannot only rely on the most-frequent-word lists and have thus instead used the full-length dictionaries.

Again, words common in all the three languages and words common in either of the two language pairs were excluded. For example, the word "*gun*" (English: weapon, Hindi: character/properties/competence/talent, Bengali: multiplication) was deleted from all three dictionaries as it is common and thus non-discriminative. Another example is the word "*din*" which is common in English (loud) and Hindi (day) dictionaries, and therefore removed. The Hindi-Bengali dictionary pair was not analysed because there are huge numbers of lexical overlaps between these two languages.

Words that cannot be found in any of these dictionaries are labelled as `undef` and passed for labelling to the subsequent module, which can consider language tags of the contextual words. This SVM-based machine learning technique is described next.

4.3. *SVM-Based Word-Language Detection*

Word-level language detection from code-mixed text can be defined as a classification problem. Support Vector Machines (SVM) were chosen for the experiment (Joachims, 1999; Joachims, 2008). The reason behind choosing SVM is that it currently is the best performing machine learning technique across multiple domains and for many tasks, including language identification (Baldwin and Lui, 2010).

For the present system, the SVM implementation in Weka (Waikato Environment for Knowledge Analysis) version 3.6.10 (Hall *et al.*, 2009) was used with default parameters. This is a linear kernel SVM, trained by Sequential Minimal Optimization, SMO (Keerthi *et al.*, 2001). The SVM classifier was trained on the following features: the n-gram list was used as a dictionary, with normalized weights for each n-gram; in addition, language specific dictionaries were used, with the MED-based weights and word context information. The details of each feature computation for the Weka-based Attribute-Relation File Format (ARFF) file creation is described below.

N-gram with weights

N-gram weight features were implemented using the bag-of-words principle. Suppose that we after pruning have n unique n-grams for the English-Hindi language pair. Then we will have n unique features. Now assume, for example, that '**IN**' is the i^{th} bi-gram in the list. In a given word w (e.g., *painting*), a particular n-gram occurs k times (twice for '**IN**' in *painting*). Then if the pre-calculated weight of the n-gram '**IN**' is ϕ_w^i, the feature vector will look as follows: $1, 2, ..., (\phi_w^i * k), .., (n-2), (n-1), n$. For any absent n-gram, the weight is set to 0. Weighting gives 3–4% better performance than binary features.

Dictionary-based features

There are three dictionaries (English, Bengali and Hindi), so there are three binary features. The presence of a word in a specific dictionary is represented by 1 and absence in the dictionary is represented by 0.

MED-based weight

If a word is absent in all dictionaries, this feature is triggered. For these *out-of-vocabulary* (OOV) words, the Minimum Edit Distance measure is calculated for each language and used as a feature, choosing the lowest distance measure as feature value. To make this search less complex, radix sort, binary search and hash map techniques were incorporated.

Word context information

A 7-word window feature (i.e., including ± 3 words around the focus word) was used to incorporate contextual information. Surface-word forms for the previous three words and their language tags along with the following three words were considered as binary features. For each word there is a unique word dictionary pre-compiled from all the corpora for both language pairs, and only three features were added for language tags. [2]

5. Experiments and Performance

A simple dictionary-based method was used as baseline, hypothesising that each text is bilingual with English as the base language. An English dictionary was used to identify each word in the text and the undefined words were marked either as Hindi or Bengali based on the corpus choice. In a real-world setting, location information could be extracted from the social media and the second language could be assumed to be the local language. For both the language pairs, the baseline performance is below 40% (38.0% and 35.5% F_1-score for English-Hindi and English-Bengali, respectively), which gives a clear indication of the difficulty.

2. An implementation detail: WEKA's SVM only takes numeral input, so instead of the actual words we use precompiled word-IDs.

System		Precision		Recall		F_1-Score	
		HND	BNG	HND	BNG	HND	BNG
N-Gram Pruning		70.12%	69.51%	48.32%	46.01%	57.21%	55.37%
+ Dictionary		82.37%	77.69%	51.03%	52.21%	63.02%	62.45%
SVM	Word Context	72.01%	74.33%	50.80%	48.55%	59.57%	58.74%
	+ N-Gram Weight	89.36%	86.83%	58.01%	56.03%	70.35%	68.11%
	+ Dictionary + MED	90.84%	87.14%	65.37%	60.22%	76.03%	74.35%

Table 5. *System performance for language detection from code-mixed text.*

5.1. *Evaluation of the Basic System Set-Up*

To understand the effect of each feature and module, experiments were carried out at various levels. The n-gram pruning and dictionary modules were evaluated separately, and those features were used in the SVM classification. The performance at the word-level on the test set is reported in Table 5. In addition, we run 10-fold cross-validation on the training set using SVM on both the language pairs and calculated the performance. The results then were quite a lot higher (with F_1-scores of around 98% and 96% for English-Hindi and English-Bengali, respectively), but as can be seen in the table, evaluation on the held-out test set made performance drop significantly. Hence, though using 10-fold cross-validation, the SVM certainly overfits the training data, which could be addressed by regularization and further feature selection. The n-gram pruning was an attempt at feature selection, but adding other features or filtering techniques is definitely possible.

Another possible solution would be to treat the language detection as a sequence labelling problem. In that case, the word-level language tag sequences should be trained using the best performing machine learning techniques for sequence labelling, such as Hidden Markov Models (HMMs) or Conditional Random Fields (CRFs). Barman *et al.* (2014) report such an attempt with a CRF-based approach, indicating a slight increase in accuracy. However, their results using CRF instead of SVM were non-conclusive in that the precision actually decreased for the majority tags, while recall increased for those tags, with the opposite tendencies for the minority tags.

It is also quite obvious from Table 5 that system performance on the English-Hindi language pair is constantly better than the English-Bengali pair. It is not totally clear why this is the case, but one possible reason can be that in the English-Hindi pair there are fewer cases of code-mixing and that they are less complex. We have not performed a separate evaluation for the formal and informal data.

System	Precision		Recall		F_1-Score	
	HND	BNG	HND	BNG	HND	BNG
Basic system	90.84%	87.14%	65.37%	60.22%	76.03%	74.35%
Post processing	94.37%	91.92%	68.04%	65.32%	79.07%	76.37%

Table 6. *Performance of the best system with and without post-processing.*

5.2. *Enhanced System with Post-Processing*

Looking at the system mistakes made on the development data, a post-processing module was designed for error correction. The most prominent errors were caused by language in continuation: Suppose that the language of the words w_n and w_{n+2} is marked by the system as l_a and that the language of the word w_{n+1} is marked as $\neg l_a$, then the post-processor's role is to restore this language to l_a. This is definitely not a linguistically correct assumption, but while working with word-level code-mixed text, this straight-forward change gives a performance boost of approximately 2–5% for both language pairs, as can be seen in Table 6, which compares the system with post-processing to the best basic system (the one shown in the last line of Table 5, i.e., SVM with word context, n-gram weight, dictionary and MED).

There are also a few errors on language boundary detection, but to post-fix those we would need to add language-specific orthographic knowledge.

5.3. *Discussion*

Social media text code-mixing in Eurasian languages is a new problem, and needs more efforts to be fully understood and solved. This linguistic phenomenon has many peculiar characteristics, for example:

[14] *addaing*

[15] *jugading*

[16] *frustu* (meaning: being frustated)

It is hard to define the language of these words, but they could be described as being examples of "*Engali*" and "*Engdi*", respectively, along the lines of Benglish and Hinglish. That is, the root forms of the words are from English, but with suffixes coming from Bengali and Hindi (see also the end of Section 3.1 and the examples in the upper part of Table 3).

Another difficult situation is reduplication, which is very frequent in South-East Asian languages (e.g., as shown by the '*majhe majhe*' construction in Example 5). English also has some reduplication (e.g., 'bye-bye'), but the phenomenon is a lot

less prominent. The social media users are influenced by the languages in their own geospaces, so reduplication is quite common in South-East Asian code-mixed text. The users in these regions are also very generative in terms of reduplication and give birth to new reduplication situations, that are not common (or even valid) in any of the Indian languages, nor in English. For example:

[17] *affair taffair*

All these phenomena contribute to complicating the language identification issue, and from the performance report and error analysis it is clear that more research efforts are needed to solve the language detection problem in the context of social media text and code-mixing. The performance of the proposed systems has only reached F_1-scores in the region of 75–80%, which is far from what would be required in order to use these techniques in a real-life setting. It is also difficult to compare the results reported here to those obtained in other media and for other types of data: while previous work on speech mainly has been on artificially generated data, previous work on text has mainly been on language identification in longer documents and at the document level, even when evidence has been collected at word level. Longer documents tend to have fewer code-switching points.

The code-mixing addressed here is more difficult and novel, and the few closely related efforts cannot be directly compared to either: the multi-lingual Twitter-setting addressed by Voss *et al.* (2014) might be closest to our work, but their results were hurt by very low precision for Moroccan Arabic, possibly since they only used a Maximum Entropy classifier to identify languages. The solution used by Carter (2012) is based on Twitter-specific priors, while the approach by Nguyen and Doğruöz (2013) utilizes language-specific dictionaries (just as our approach does), making a comparison across languages somewhat unfair. The idea introduced by Lignos and Marcus (2013), to only use the ratio of the word probability, would potentially be easier to compare across languages.

Our work also substantially differs from Nguyen and Doğruöz (2013) and Lignos and Marcus (2013) by addressing a multi-lingual setting, while their work is strictly bilingual (with the first authors making the assumption that words from other languages — English — appearing in the messages could be assumed to belong to the dominating language, i.e., Dutch in their case). Further, even though they also work on chat data, Nguyen and Doğruöz (2013) mainly investigated utterance (post) level classification, and hence give no actual word-level baseline, but just state that 83% of the posts are monolingual. 2.71% of their unique tokens are multi-lingual, while in our case it is 8.25%. Nguyen and Doğruöz have gratefully made their data available. Testing our system on it gives a slightly increased accuracy compared to their results (by 0.99%).

For a partial remedy to the problem of comparing code-mixed corpora from different types of text, genres, and language pairs, see Gambäck and Das (2014) where we introduce and discuss a Code-Mixing Index specifically designed to make this comparison possible. The Code-Mixing Index is based on information about the frequency of words from the most common language in each single utterance, but taken on average over all utterances.

6. Conclusion

Language evolution is arguably a difficult problem to solve and is highly interdisciplinary in nature (Christiansen and Kirby, 2003; de Boer and Zuidema, 2010). The social media revolution has added a new dimension to language evolution, with the borders of society fading, and the mixing of languages and cultures increasing.

The paper has presented an initial study on the detection of code-mixing in the context of social media texts. This is a quite complex language identification task which has to be carried out at the word-level, since each message and each single sentence can contain text and words in several languages. The experiments described in here have focused on code-mixing only in Facebook posts written in the language pairs English-Hindi and English-Bengali, from a corpus collected and annotated as part of the present work. This is on-going work and the performance of the proposed systems has only reached 75–80%, which is far from what would be required in order to use these techniques in a real-life setting. However, the work is novel in terms of problem definition and in terms of resource creation.

In the future, it would be reasonable to experiment with other languages and other types of social media text, such as tweets (Carter, 2012; Solorio *et al.*, 2014). Although Facebook posts tend to be short, they are commonly not as short as tweets, which have a strict length limitation (to 140 characters). It would be interesting to investigate whether this restriction induces more or less code-mixing in tweets (as compared to Facebook posts), and whether the reduced size of the context makes language identification even harder.

The language identification system described here mainly uses standard techniques such as character n-grams, dictionaries and SVM-classifiers. Incorporating other techniques and information sources are obvious targets for future work. In particular, to look at other machine learning methods, for example, to use a sequence learning method such as Conditional Random Fields (Nguyen and Doğruöz, 2013; Barman *et al.*, 2014) to capture patterns of sequences containing code switching, or to use combinations (ensembles) of different types of learners.

Acknowledgements

Thanks to Dong Nguyen (University of Twente, The Netherlands) and Seza Doğruöz (Tilburg University, The Netherlands) for making their data set available, and to Utsab Barman (Dublin City University, Ireland) for helping us with the corpus collection and annotation.

Special thanks to Sandrine Henry and several anonymous reviewers for comments that over time have substantially improved the paper.

7. References

Ahmed B., Cha S.-H., Tappert C., "Language Identification from Text Using N-gram Based Cumulative Frequency Addition", *Proceedings of Student/Faculty Research Day*, School of Computer Science and Information Systems, Pace University, New York, USA, p. 12:1-12:8, 2004.

Ahmed B. U., Detection of Foreign Words and Names in Written Text, PhD Thesis, School of Computer Science and Information Systems, Pace University, New York, USA, 2005.

Alex B., Automatic Detection of English Inclusions in Mixed-lingual Data with an Application to Parsing, PhD Thesis, School of Informatics, University of Edinburgh, Edinburgh, UK, 2008.

Ali I., Mahmood Aslam T., "Frequency of Learned Words of English as a Marker of Gender Identity in SMS Language in Pakistan", *Journal of Elementary Education*, vol. 22, n° 2, p. 45-55, 2012.

Andersen G., "Semi-automatic approaches to Anglicism detection in Norwegian corpus data", *in* C. Furiassi, V. Pulcini, F. R. González (eds), *The Anglicization of European lexis*, John Benjamins, p. 111-130, 2012.

Auer P., *Bilingual Conversation*, John Benjamins, 1984.

Auer P., "From codeswitching via language mixing to fused lects: Toward a dynamic typology of bilingual speech", *International Journal of Bilingualism*, vol. 3, n° 4, p. 309-332, 1999.

Baldwin T., "Lexical normalisation dictionary", 2012. http://www.csse.unimelb.edu.au/~tim/etc/emnlp2012-lexnorm.tgz.

Baldwin T., Lui M., "Language Identification: The Long and the Short of the Matter", *Proceedings of the 2010 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, ACL, Los Angeles, California, p. 229-237, June, 2010.

Barman U., Das A., Wagner J., Foster J., "Code Mixing: A Challenge for Language Identification in the Language of Social Media", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, ACL, Doha, Qatar, p. 13-23, October, 2014. 1st Workshop on Computational Approaches to Code Switching.

Beesley K. R., "Language Identifier: A Computer Program for Automatic Natural-Language Identification of On-line Text", *Proceedings of the 29th Annual Conference of the American Translators Association*, Medford, New Jersey, p. 47-54, 1988.

Biśvās Ś., *Samsad Bengali-English dictionary*, 3 edn, Sahitya Samsad, Calcutta, India, 2000.

Bock Z., "Cyber socialising: Emerging genres and registers of intimacy among young South African students", *Language Matters: Studies in the Languages of Africa*, vol. 44, n° 2, p. 68-91, 2013.

Bontcheva K., Cunningham H., Roberts I., Roberts A., Tablan V., Aswani N., Gorrell G., "GATE Teamware: a web-based, collaborative text annotation framework", *Language Resources and Evaluation*, vol. 47, n° 4, p. 1007-1029, December, 2013.

Bullock B. E., Hinrichs L., Toribio A. J., "World Englishes, code-switching, and convergence", *in* M. Filppula, J. Klemola, D. Sharma (eds), *The Oxford Handbook of World Englishes*, Oxford University Press, Oxford, England, 2014. Forthcoming. Online publication: March 2014.

Carter S., Exploration and Exploitation of Multilingual Data for Statistical Machine Translation, PhD Thesis, University of Amsterdam, Informatics Institute, Amsterdam, The Netherlands, December, 2012.

Carter S., Weerkamp W., Tsagkias M., "Microblog language identification: overcoming the limitations of short, unedited and idiomatic text", *Language Resources and Evaluation*, vol. 47, n° 1, p. 195-215, March, 2013. Special Issue on Analysis of short texts on the Web.

Cavnar W. D., Trenkle J. M., "N-Gram-Based Text Categorization", *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, UNLV Publications/Reprographics, Las Vegas, Nevada, p. 161-175, April, 1994.

Center for Indian Language Technology, "Hindi Wordnet: A Lexical Database for Hindi", January, 2013. http://www.cfilt.iitb.ac.in/wordnet/webhwn/.

Chan J. Y., Cao H., Ching P., Lee T., "Automatic Recognition of Cantonese-English Code-Mixing Speech", *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 14, n° 3, p. 281-304, 2009.

Christiansen M. H., Kirby S., "Language Evolution: The Hardest Problem in Science?", *in* M. H. Christiansen, S. Kirby (eds), *Language Evolution*, Oxford University Press, Oxford, England, p. 1-15, 2003.

Damashek M., "Gauging Similarity with n-Grams: Language-Independent Categorization of Text", *Science*, vol. 267, n° 5199, p. 843-848, 1995.

Das A., Gambäck B., "Identifying Languages at the Word Level in Code-Mixed Indian Social Media Text", *Proceedings of the 11th International Conference on Natural Language Processing*, Goa, India, p. 169-178, December, 2014.

Das A., Saikh T., Mondal T., Ekbal A., Bandyopadhyay S., "English to Indian Languages Machine Transliteration System at NEWS 2010", *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL, Uppsala, Sweden, p. 71-75, July, 2010. 2nd Named Entities Workshop.

de Boer B., Zuidema W., "Models of Language Evolution: Does the Math Add Up?", *Proceedings of the 8th International Conference on the Evolution of Language*, Utrecht, the Netherlands, p. 1-10, April, 2010. Workshop on Models of Language Evolution.

Dewaele J.-M., "The emotional weight of *I love you* in multilinguals' languages", *Journal of Pragmatics*, vol. 40, n° 10, p. 1753-1780, October, 2008.

Dewaele J.-M., *Emotions in Multiple Languages*, Palgrave Macmillan, 2010.

Digital South Asia Library, "Digital Dictionaries of South Asia — Sailendra Biswas: SAMSAD BENGALI-ENGLISH DICTIONARY", February, 2006. http://dsal.uchicago.edu/dictionaries/biswas-bengali/.

Dunning T., Statistical Identification of Language, Technical report, Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico, March, 1994.

Fischer E., "Language communities of Twitter", October, 2011. http://www.flickr.com/photos/walkingsf/6277163176/in/photostream/.

Gafaranga J., Torras M.-C., "Interactional otherness: Towards a redefinition of codeswitching", *International Journal of Bilingualism*, vol. 6, n° 1, p. 1-22, 2002.

Gambäck B., Das A., "On Measuring the Complexity of Code-Mixing", *Proceedings of the 11th International Conference on Natural Language Processing*, Goa, India, p. 1-7, December, 2014. 1st Workshop on Language Technologies for Indian Social Media.

Gold E. M., "Language Identification in the Limit", *Information and Control*, vol. 10, n° 5, p. 447-474, 1967.

Gottron T., Lipka N., "A Comparison of Language Identification Approaches on Short, Query-Style Texts", *Advances in Information Retrieval: 32nd European Conference on IR Research, Proceedings*, Springer, Milton Keynes, UK, p. 611-614, March, 2010.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H., "The WEKA Data Mining Software: An Update", *ACM SIGKDD Explorations Newsletter*, vol. 11, n° 1, p. 10-18, November, 2009.

Han B., Cook P., Baldwin T., "Automatically Constructing a Normalisation Dictionary for Microblogs", *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, ACL, Jeju Island, Korea, p. 421-432, July, 2012.

Hidayat T., *"An Analysis of Code Switching Used by Facebookers (a Case Study in a Social Network Site)"*, BA Thesis, English Education Study Program, College of Teaching and Education (STKIP), Bandung, Indonesia, October, 2012.

Joachims T., "Making Large-Scale Support Vector Machine Learning Practical", *in* B. Schölkopf, C. J. Burges, A. J. Smola (eds), *Advances in Kernel Methods: Support Vector Learning*, MIT Press, Cambridge, Massachusetts, chapter 11, p. 169-184, 1999.

Joachims T., "SVM^{struct}: Support Vector Machine for Complex Outputs", August, 2008. http://www.cs.cornell.edu/people/tj/svm_light/svm_struct.html.

Johar M. M. B., "The Effect of Emotional Arousal on Code-Switching in Social Network-Mediated Micro-Blogging", *Vernaculum*, n° 2, p. 21-29, 2011.

Joshi A. K., "Processing of Sentences with Intra-sentential Code-switching", *Proceedings of the 9th International Conference on Computational Linguistics*, ACL, Prague, Czechoslovakia, p. 145-150, July, 1982.

Keerthi S. S., Shevade S. K., Bhattacharyya C., Murthy K. R. K., "Improvements to Platt's SMO Algorithm for SVM Classifier Design", *Neural Computation*, vol. 13, n° 3, p. 637-649, March, 2001.

Khapra M. M., Joshi S., Ramanathan A., Visweswariah K., "Offering Language Based Services on Social Media by Identifying User's Preferred Language(s) from Romanized Text", *Proceedings of the 22nd International World Wide Web Conference*, vol. Companion, Rio de Janeiro, Brazil, p. 71-72, May, 2013.

King B., Abney S., "Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods", *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, Atlanta, Georgia, p. 1110-1119, June, 2013.

Kishi Adelia N., *"Investigating the Types and Functions of Code Switching on Twitter's Tweets by Male and Female Students of English Department, Binus University"*, BA Thesis, School of English Literature, Binus University, Jakarta, Indonesia, 2012.

Levenshtein V. I., "Binary Codes Capable of Correcting Deletions, Insertions and Reversals", *Soviet Physics Doklady*, vol. 10, n° 8, p. 707-710, February, 1966.

Li D. C. S., "Cantonese-English code-switching research in Hong Kong: a Y2K review", *World Englishes*, vol. 19, n° 3, p. 305-322, November, 2000.

Lignos C., Marcus M., "Toward Web-scale Analysis of Codeswitching", *87th Annual Meeting of the Linguistic Society of America*, Boston, Massachusetts, January, 2013. Poster.

Lui M., Baldwin T., "Accurate Language Identification of Twitter Messages", *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, ACL, Göteborg, Sweden, p. 17-25, April, 2014. 5th Workshop on Language Analysis for Social Media.

Lui M., Lau J. H., Baldwin T., "Automatic Detection and Language Identification of Multilingual Documents", *Transactions of the Association for Computional Linguistics*, vol. 2, p. 27-40, February, 2014.

McNamee P., "Language Identification: A Solved Problem Suitable for Undergraduate Instruction", *Journal of Computing Sciences in Colleges*, vol. 20, n° 3, p. 94-101, February, 2005.

Meganathan R., Language policy in education and the role of English in India: From library language to language of empowerment, Dreams and Realities: Developing Countries and the English Language n° 4, British Council, London, England, 2011.

Muysken P., "Code-switching and grammatical theory", *in* L. Milroy, P. Muysken (eds), *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, Cambridge University Press, Cambridge, England, p. 177-198, 1995.

Muysken P., *Bilingual speech: A typology of code-mixing*, Cambridge University Press, Cambridge, England, 2000.

Narayan D., Chakrabarti D., Pande P., Bhattacharyya P., "An Experience in Building the Indo WordNet — a WordNet for Hindi", *Proceedings of the 1st International Conference on Global WordNet*, Mysore, India, January, 2002.

Negrón Goldbarg R., "Spanish-English Codeswitching in Email Communication", *Language@Internet*, vol. 6, p. article 3, February, 2009.

Nguyen D., Doğruöz A. S., "Word Level Language Identification in Online Multilingual Communication", *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, ACL, Seattle, Washington, p. 857-862, October, 2013.

Peng N., Wang Y., Dredze M., "Learning Polylingual Topic Models from Code-Switched Social Media Documents", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 2, ACL, Baltimore, Maryland, p. 674-679, June, 2014.

Prager J. M., "Linguini: Language Identification for Multilingual Documents", *Proceedings of the 32nd Hawaii International Conference on Systems Sciences*, IEEE, Maui, Hawaii, p. 1-11, January, 1997.

Řehůřek R., Kolkus M., "Language Identification on the Web: Extending the Dictionary Method", *in* A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing: Proceedings of the 10th International Conference*, n° 5449 in *Lecture Notes in Computer Science*, Springer-Verlag, Mexico City, Mexico, p. 357-368, March, 2009.

Rodrigues P., Processing Highly Variant Language Using Incremental Model Selection, PhD Thesis, Indiana University, Dept. of Linguistics, Bloomington, Indiana, February, 2012.

Rodrigues P., Kübler S., "Part of Speech Tagging Bilingual Speech Transcripts with Intrasentential Model Switching", *Papers from the AAAI Spring Symposium on Analyzing Microtext*, AAAI, Stanford University, California, p. 56-65, March, 2013.

Rosner M., Farrugia P.-J., "A Tagging Algorithm for Mixed Language Identification in a Noisy Domain", *Proceedings of the 8th Annual INTERSPEECH Conference*, vol. 3, ISCA, Antwerp, Belgium, p. 1941-1944, August, 2007.

San H. K., *"Chinese-English Code-switching in Blogs by Macao Young People"*, MSc Thesis, Applied Linguistics, University of Edinburgh, Edinburgh, Scotland, August, 2009.

Schroeder S., "Half of Messages on Twitter Aren't in English [STATS]", February, 2010. http://mashable.com/2010/02/24/half-messages-twitter-english/.

Shafie L. A., Nayan S., "Languages, Code-Switching Practice and Primary Functions of Facebook among University Students", *Study in English Language Teaching*, vol. 1, nº 1, p. 187-199, February, 2013.

Singh A. K., Gorla J., "Identification of Languages and Encodings in a Multilingual Document", *Proceedings of the 3rd Workshop on Building and Exploring Web Corpora*, Presses universitaires de Louvain, Louvain-la-Neuve, Belgium, p. 95-108, September, 2007.

Solorio T., Blair E., Maharjan S., Bethard S., Diab M., Gohneim M., Hawwari A., AlGhamdi F., Hirschberg J., Chang A., Fung P., "Overview for the First Shared Task on Language Identification in Code-Switched Data", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, ACL, Doha, Qatar, p. 62-72, October, 2014. 1st Workshop on Computational Approaches to Code Switching.

Solorio T., Liu Y., "Learning to Predict Code-Switching Points", *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, ACL, Honolulu, Hawaii, p. 973-981, October, 2008a.

Solorio T., Liu Y., "Part-of-Speech Tagging for English-Spanish Code-Switched Text", *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, ACL, Honolulu, Hawaii, p. 1051-1060, October, 2008b.

Solorio T., Sherman M., Liu Y., Bedore L. M., Peña E. D., Iglesias A., "Analyzing language samples of Spanish-English bilingual children for the automated prediction of language dominance", *Natural Language Engineering*, vol. 17, nº 3, p. 367-395, July, 2011.

Sotillo S., "*Ehhhh utede hacen plane sin mi???:@ im feeling left out:(* Form, Function and Type of Code Switching in SMS Texting", *ICAME 33 Corpora at the centre and crossroads of English linguistics*, Katholieke Universiteit Leuven, Leuven, Belgium, p. 309-310, June, 2012.

Teahan W. J., "Text classification and segmentation using minimum cross-entropy", *Proceedings of the 6th International Conference on Computer-Assisted Information Retrieval (Recherche d'Information Assistée par Ordinateur, RIAO 2000)*, Paris, France, p. 943-961, April, 2000.

van Noord G., "TextCat", 1997. http://odur.let.rug.nl/~vannoord/TextCat/.

Voss C., Tratz S., Laoudi J., Briesch D., "Finding Romanized Arabic Dialect in Code-Mixed Tweets", *Proceedings of the 9th International Conference on Language Resources and Evaluation*, ELRA, Reykjavík, Iceland, p. 188-199, May, 2014.

Weiner J., Vu N. T., Telaar D., Metze F., Schultz T., Lyu D.-C., Chng E.-S., Li H., "Integration of language identification into a recognition system for spoken conversations containing code-switches", *Proceedings of the 3rd Workshop on Spoken Language Technologies for Under-resourced Languages*, Cape Town, South Africa, p. 76-79, May, 2012.

Xochitiotzi Zarate A. L., "Code-mixing in Text Messages: Communication Among University Students", *Memorias del XI Encuentro Nacional de Estudios en Lenguas*, Universidad Autónoma de Tlaxcala, Tlaxcala de Xicohtencatl, Mexico, p. 500-506, 2010.

Yamaguchi H., Tanaka-Ishii K., "Text segmentation by language using minimum description length", *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, vol. 1, ACL, Jeju, Korea, p. 969-978, July, 2012.

Notes de lecture

Rubrique préparée par Denis Maurel

Université François Rabelais Tours, LI (Laboratoire d'informatique)

Agnès TUTIN, Francis GROSSMANN. L'écrit scientifique, du lexique au discours : autour de Scientext. *Presses universitaires de Rennes*. 2013. 232 pages. ISBN 978-2-7535-2846-8.

Lu par **Nadine LUCAS**

GREYC - CNRS UMR 6072 – Université de Caen Basse-Normandie

L'étude de l'écrit scientifique vise en particulier à mieux cerner ce qui fait la spécificité du langage scientifique et à comparer les rhétoriques disciplinaires, étudiées à travers un corpus contemporain sur les sciences humaines et sociales mais aussi les sciences expérimentales (médecine) et les sciences appliquées. Au centre des préoccupations, il y a la manière dont sont mises en scène, dans le lexique et les discours, les procédures de découverte et de validation des connaissances. L'ouvrage s'appuie sur un ensemble de recherches effectuées à partir du corpus Scientext, corpus en français et en anglais développé par plusieurs équipes de chercheurs sous la responsabilité du LIDILEM à Grenoble 3. Deux problématiques sont privilégiées, le positionnement *sociolinguistique de l'auteur scientifique par rapport à ses devanciers et à ses pairs, et celle du* raisonnement*, tel qu'il apparaît dans l'argumentation ou les chaînes causales.*

Contenu et structure

Cet ouvrage collectif porte sur les retombées d'un projet de l'ANR, Scientext, achevé fin 2009. Il concerne la constitution, et l'exploitation *via* une interface informatique, d'un corpus scientifique, composé d'articles et de thèses en français, soit 219 textes consultables (4,8 millions de mots). Le sous-corpus des didacticiens (300 textes, 1,1 million de mots) est beaucoup plus restreint et non partagé, de même que le corpus exploratoire sur les évaluations de communications de colloques (502 documents, 34 857 mots). Le corpus en anglais issu de BMC est beaucoup plus conséquent : 3 381 articles (13,8 millions de mots) dans le domaine médical et biologique.

Une introduction et une bibliographie, suivie des résumés des contributions, encadrent trois parties d'inégale longueur. L'ouvrage compte dix contributions assez homogènes, huit sont issues du LIDILEM, maître d'œuvre du projet. Sur les neuf contributions centrales, la majorité aborde l'aspect lexical et explicite de l'écrit académique, en recensant dans des sous-corpus les occurrences de termes emblématiques : le verbe *voir*, le verbe *causer*, les collocations telles que *On constate* ou *résultats intéressants*. Les études sur le français dominent, deux seulement portent sur l'anglais. La constitution d'un dictionnaire dynamique est

traitée par G. Williams et C. Millon à partir d'un gros corpus de médecine en anglais. Quatre contributions ont des problématiques un peu différentes. L'une, en fin de première partie, traite de l'ingénierie du projet, deux contributions, réunies dans la deuxième partie, traitent d'applications pédagogiques, enfin une seule contribution en troisième partie traite de la titraille des articles du corpus.

L'objectif était de permettre l'outillage de l'étude de phénomènes linguistiques sur corpus par des linguistes, grâce notamment à ScienQuest, une interface simple d'interrogation. Cet objectif est atteint. Les conclusions des études de corpus reflètent l'utilité des concordanciers et la satisfaction des linguistes de pouvoir recenser les occurrences recherchées. Sur le plan de l'analyse linguistique, les conclusions montrent cependant les limites du dispositif. Elles reflètent une certaine déception, souvent explicite, par rapport aux hypothèses de constance d'une phraséologie scientifique et de représentativité des expressions recherchées. Les mots candidats sont peu fréquents et inégalement répartis entre les disciplines.

Commentaire

Voici donc un projet qui vise « à outiller la linguistique ». Sans doute faut-il se rappeler qu'il ne s'agit pas ici d'un ouvrage destiné à un public de TAL. L'exploration des outils prend du temps et ce sont donc les premiers pas d'utilisateurs presque tous novices qui sont présentés ici. Sans entrer dans le détail, nous relevons les tendances générales et les exceptions qui confirment la règle.

L'article d'A. Falaise sur le corpus et les outils est curieusement situé tout à la fin de la première partie, et donne donc tardivement la perspective des moyens mis en œuvre à l'intention des linguistes. Avec le logiciel ScienQuest, il leur propose un outil « intuitif » pour explorer leur corpus, mais cela se solde souvent par une réduction drastique des choix possibles à l'aide de menus. On notera aussi que le site de Scientext n'est pas mis en avant : on trouve son adresse[1] p. 16 de l'introduction et dans les références.

Il est vrai que Scientext ne dissocie pas clairement le corpus, le modèle et les outils. À ce titre, il ne semble pas pouvoir se prêter à d'autres problématiques que celles retenues par les premiers partenaires. Dans ScienQuest, le mode de consultation dit « sémantique » en décevra plus d'un, c'est en effet la vue par défaut, donnant les résultats de grammaires déjà construites en fonction des questions traitées par l'équipe. Le mode libre permet de construire une recherche à l'aide d'un assistant, enfin le mode avancé permet de faire des requêtes classiques avec opérateurs et distance entre termes.

Si la base Scientext représente un corpus relativement varié pour des linguistes, elle souffre de limitations d'accès, du fait des droits d'auteur : de fait, la moitié du corpus brut collecté en français n'est consultable qu'en intranet, et le corpus analysé par Syntex ne l'est pas non plus. On est encore loin de la diversité des corpus et des études menées par l'école scandinave ou des outils statistiques de l'école anglaise.

1. http://scientext.mrsh.alpes.fr.

Par ailleurs, la majorité des linguistes se font encore une piètre idée de l'informatique, jugeant par exemple que détecter des expressions discontinues est « très difficile ». En conséquence, ils ont très peu recours à la consultation en mode libre ou avancé, plus souple que les concordanciers. Les résultats de l'analyseur syntaxique Syntex de Bourigault restent sous-utilisés. Il y a donc matière à apprentissage et à progression dans ce domaine.

Bien des auteurs se sont montrés méfiants : ils se sont contentés d'ajouter des décomptes d'occurrences à leur description du français académique contemporain, sans étendre leur sous-corpus. La majorité des études portent sur des collections d'une trentaine de textes, principalement sur les sciences humaines. À l'opposé, la contribution de M.-P. Jacques fait exception en traitant de l'ensemble du corpus français disponible, à travers leurs titres et sous-titres. L'étude de Williams et Millon se réfère à une vaste collection d'articles en anglais issue du corpus médical BMC, dans le but de générer un dictionnaire.

D'un point de vue méthodologique, la majorité des études sont centrées sur l'usage du lexique à travers quelques disciplines, mesuré par la fréquence brute d'occurrences. Il est assez étonnant pour des linguistes que l'approche reste attachée au métalangage explicite, par exemple « nous nous appuyons sur Untel » pour étudier les emprunts de méthode. Mais le logocentrisme ne devrait pas heurter un public de talistes, également attaché à la valeur littérale des indices, le « sémantisme » des mots.

Du moins, l'ouvrage aura le mérite de montrer répétitivement que la valeur lexicale ne suffit pas à capter le raisonnement causal, ou le positionnement sociolinguistique, les deux problématiques qui sous-tendent les investigations. Un constat important affleure : la représentativité des termes relève du prototype sémantique et ne correspond nullement à une grande fréquence d'emploi, non plus qu'à une distribution uniforme et transdisciplinaire des termes considérés.

La bibliographie est bien développée sur l'approche linguistique logocentrique et l'approche sociolinguistique. En revanche, elle ne traite pas des aspects stylistique et pédagogique de la rhétorique scientifique, ni de la relation entre l'illustration et le texte. Sans surprise, elle n'établit aucun lien non plus entre l'étude des textes scientifiques et les travaux de fouille de textes associant linguistes et informaticiens.

Pour les informaticiens linguistes, malheureusement, la base de textes Scientext est peu exploitable. Elle est trop restreinte, à l'heure des *big data*, à la fois en nombre de textes et de disciplines représentées. L'impossibilité d'accès direct au corpus français, brut ou annoté est rédhibitoire pour l'exploration occasionnelle ; le seul corpus téléchargeable est celui de BMC en anglais et il était déjà disponible.

La plate-forme Scientext peut cependant aider les talistes lexicologues à explorer semi-manuellement le corpus proposé, par exemple pour détecter des patrons syntaxiques associés à des formules d'usage, la « phraséologie ». Il est également possible d'avoir accès sur demande au corpus brut.

Jesse TSENG. Prépositions et postpositions : approches typologiques et formelles. *Hermès-Lavoisier*. 2013. 253 pages. ISBN 978-2-7462-4518-1.

Lu par **Marie-Hélène LAY**

Laboratoire FoReLL (EA3616) – Université de Poitiers

Jesse Tseng propose ici un ouvrage consacré à l'étude des prépositions, présentées comme une catégorie lexicale « majeure » pour laquelle la question de la faiblesse grammaticale se pose tout particulièrement. De taille restreinte, cette catégorie présente des éléments caractérisés par la multitude de leurs emplois (due à la polysémie et à des phénomènes de grammaticalisation) et par le fait qu'ils entrent dans des constructions « contraintes » (syntaxiquement déterminées ou lexicalement figées). Souvent considérée comme inerte sur le plan morphologique (absence de productivité flexionnelle et dérivationnelle), cette catégorie peut présenter, par ailleurs, des comportements morpho-phonologiques complexes (contraction). Les six contributions de ce volume sont des contributions longues (de 25 à 40 pages environ) abordant la question des prépositions faibles pour six langues différentes. Les faits étudiés sont variés, comme le sont les perspectives descriptives et analytiques abordées.

Le chapitre 1 offre une présentation détaillée de l'inventaire des diverses formes susceptibles d'être employées comme des prépositions en tswana. Elles constituent une classe hétérogène du point de vue phonologique, morphologique et syntaxique comme du point de vue de leurs emplois, ce qui conduit l'auteur à les analyser en comparaison avec d'autres connecteurs et joncteurs. La contribution traite des éléments situés à la marge gauche des constituants nominaux qui informent sur les relations syntaxiques et sémantiques que ce constituant entretient avec le reste de la phrase. Ils sont très largement analysés comme des préfixes du nom par la tradition, bien qu'un petit nombre d'entre eux (ce qui est une propriété des langues subsahariennes) puisse fonctionner comme des éléments autonomes indécomposables. Leur identification en tant que « mot » repose sur la propagation postlexicale du ton haut et sur la possibilité ou l'impossibilité de l'apparition d'un abaissement du registre haut, le *downstep* étant la manifestation de la démarcation lexicale. *Le*, *ka*, *ke*, d'une part, *-a*, *go -ng* d'autre part et enfin *ko*, *fa*, *mo* sont successivement étudiés dans leurs emplois comme prépositions, connecteurs et joncteurs liés au constituant ou pas.

Le chapitre 2 présente une étude contrastive des contractions préposition et article en français et en allemand. L'analyse proposée ici avance que les contractions sont dues dans les deux langues à des composantes différentes de la grammaire. La contraction en français s'effectue au niveau phono-morphologique et produit des prépositions fléchies appartenant au lexique présyntaxique, elle est obligatoire et exclut de ce fait certaines formes de coordination distante sans reprise de la préposition. Elle est déterminée par le genre et le nombre du mot (masculin singulier ou pluriel) qui suit, mais ne se produit pas s'il y a élision de la voyelle de l'article (*à l'homme*). En revanche en allemand, ces formes peuvent apparaître contractées ou non, dans des contextes comparables mais avec des interprétations distinctes, selon

qu'il s'agit d'une lecture de définitude pragmatique ou de définitude sémantique. Ce phénomène postsyntaxique est déterminé lexicalement et pas phonologiquement. Par ailleurs, aussi bien la préposition que le déterminant gardent une indépendance syntaxique malgré la forme morphologique qui amalgame les deux, ce qui ne serait pas le cas en français, les amalgames n'occupant alors qu'une position syntaxique.

Le chapitre suivant se penche sur le statut morphosyntaxique délicat des prépositions : elles apparaissent sous une forme préfixée, sans être pour autant des marqueurs casuels mais bien plutôt des prépositions présentant une déficience phonologique leur interdisant de se projeter en catégorie autonome. L'analyse de la distribution des six prépositions faibles du kabyle *f* (sur), *s* (avec, instrument), *o* (comitatif), *g* (dans), *i* (datif) et *n* (génitif) amène l'auteur à conclure qu'elles sont la tête du syntagme lorsqu'elles précèdent un nom à l'état d'annexion ; elles font alors partie du domaine phonologique de leur complément nominal et ne sont pas des marqueurs casuels au sens strict : elles n'introduisent pas de complément phrastique régi dans ces contextes. Des considérations sur le gabarit des complémenteurs amènent à la conclusion que les prépositions faibles affixées ne peuvent pas être visibles en syntaxe. Leur affixation dans ce contexte est un processus de composition, et non de concaténation syntaxique, ce qui permet en outre d'expliquer les restrictions portant sur le redoublement des prépositions faibles dans les constructions à long mouvement *wh*.

Le chapitre 4 concerne un aspect particulier des postpositions du coréen afin de rendre compte des empilements possibles de plusieurs éléments portant sur le même constituant. Rompant avec une tradition les envisageant comme des cas, l'auteur les traite ici comme des enclitiques avec un statut syntaxique de tête faible. Ils sont formalisés en HPSG et traités comme relevant de trois sous-catégories : marqueurs syntaxiques cas et relations syntaxiques. *-i*, *-leul- -eun*, *-e* sont étudiés. Elles s'attachent à un item lexical comme des suffixes dont elles ne partagent pourtant pas les comportements. Elles se comportent comme des clitiques qui se combinent avec un syntagme et s'attachent en phonologie au dernier mot. Elles sont analysées comme des têtes faibles qui prennent pour complément le syntagme dominant l'hôte phonologique dont elles héritent les propriétés syntaxiques. Cette analyse permet l'économie d'une distinction entre deux homonymes « postposition casuelle / non-casuelle ». Elle simplifie aussi la description des empilements, explicitant les contraintes sur l'ordre et la cooccurrence.

La contribution du chapitre 5 propose une classification et une analyse formelle de prépositions de certains parlers kurdes. Il s'agit des prépositions qui présentent deux allomorphes à l'exclusion des prépositions composées et des prépositions nominales. La *forme simple* de la préposition se combine avec des compléments de forme pleine (forte) alors que sa *forme absolue* n'accepte que les compléments pronominaux de forme faible. Cette dernière est étudiée plus en détail, afin de rendre plus précisément compte du mode de la réalisation de ces compléments : les morphèmes personnels liés peuvent être divisés en deux classes selon leurs propriétés de placement, clitique et désinence personnelle sur le verbe, apparaissant en distribution complémentaire (en corrélation avec les aspects transitif et/ou intransitif des verbes et le temps employé). Un traitement formel (HPSG) de

l'alternance entre les formes simples et absolues est ensuite proposé. Il repose d'une part sur une classification des propositions intégrant deux dimensions (le caractère nominal ou non de la préposition et la réalisation argumentale – constituant ou affixe), et d'autre part sur les informations contenues dans les entrées lexicales respectives des membres de chaque classe.

Le chapitre 6 présente une étude comparée des deux formes *pe* et *a* qui, outre des emplois prépositionnels comparables pour partie à ceux de *à* et *sur* en français, marquent, sous certaines conditions, l'objet direct d'un verbe transitif : elles projettent alors des groupes nominaux dont la nature est celle de leur complément. Leur apparition peut être corrélée avec le type sémantique de l'objet et avec les propriétés des structures dites à incorporation : ils ont nécessairement une dénotation de type individu ne pouvant donc pas être incorporés sémantiquement et sont exclus des objets directs ayant une dénotation de type propriété. Par ailleurs, un objet direct non incorporé peut être interprété comme topicalisé. Deux formalisations sont alors suggérées, les analysant soit comme des marques de topicalisation, soit comme des marques de cas fort. Les facteurs déterminant la présence ou l'absence de *pe* et *a* sont organisés en une échelle hiérarchique formée de trois paramètres ayant des valeurs graduelles : le caractère animé de l'objet direct, son caractère spécifique et sa topicalisation. Dans une perspective comparative, ce sont ces valeurs qui fondent la différence de l'emploi de *pe* et de *a*, seul l'espagnol étant sensible à la topicalisation.

Ces travaux peuvent sembler ardus à un étudiant de TAL, mais ils permettent d'aborder de façon concrète (par la diversité des langues étudiées) un certain nombre de problèmes centraux pour la linguistique générale : par exemple l'imbrication complexe des « niveaux » de description que l'on tend trop souvent à vouloir dissocier strictement : il y est régulièrement question de problèmes phono-morphologiques, morphosyntaxiques voire phono-syntaxiques.

Philipp CIMIANO, Christina UNGER, John McCRAE. Ontology-Based Interpretation of Natural Language. *Morgan & Claypool publishers*. 2014. 155 pages. ISBN 978-1-6084-5989-6.

Lu par **François LÉVY**

Université Paris 13 – LIPN

Ce livre propose une architecture de TAL modulaire, fondée sur les standards du Web et des données ouvertes. L'analyse syntaxique et la représentation sémantique sont menées en parallèle. Les données pour cela sont compilées à partir d'un lexique disposant d'une large palette de notations linguistiques et de références à une ontologie. Le lexique et l'ontologie utilisent des formats standard du Web, ce qui les rend échangeables et réutilisables.

Le livre publié par nos trois collègues de Bielefeld est issu d'un cours à l'ESSLI et dans leur université. Il réussit en cent cinquante pages un exposé clair et pédagogique de l'architecture de TAL qu'ils proposent. C'est le terme *interpretation*

employé dans le titre qui m'a donné envie d'y regarder de plus près, en ce qu'il évoque une conception non compositionnelle. Effectivement, le lien entre texte et représentation sémantique ne se limite pas à un décalque de l'analyse syntaxique. Il repose sur une conception du lexique à laquelle les auteurs ont consacré beaucoup d'efforts, et qui est intéressante. Je schématise la conception de l'ensemble avant de proposer une analyse plus personnelle de ses perspectives.

Formalismes et calculs

Au départ, il y a un texte et un ensemble formalisé de connaissances *a priori* sur le monde. À l'arrivée, des connaissances formalisées supplémentaires apportées par le texte. Les connaissances sur le monde sont décrites dans une ontologie. Le chapitre 2 rappelle les principes et les axiomes qui sous-tendent OWL DL et OWL2 DL et leur correspondance avec une partie de la logique du premier ordre (LPO). L'interprétation utilise une grammaire d'arbres adjoints lexicalisée (LTAG) pour l'analyse syntaxique. Le chapitre 3 explique le couplage de cette grammaire et du formalisme sémantique des DUDES[2] qui est pour l'essentiel celui de la théorie des représentations discursives (DRT), augmenté pour les besoins du couplage. Très schématiquement, chaque arbre syntaxique élémentaire représente un mot ou une expression avec les places syntaxiques qui seront utilisées pour sa combinaison avec les autres arbres de la phrase. De plus, chaque arbre élémentaire est associé à une DRS qui en constitue la représentation sémantique : le vocabulaire est celui de l'ontologie (dans une formulation en LPO). Les *paires sélectives* qui augmentent la DRS associent chacune une place syntaxique de l'arbre et une variable de l'univers de la DRS ; de plus, la *variable principale* est celle qui est marquée pour une possible unification. Grâce à ce couplage, les opérations de combinaison des DUDES sont définies en parallèle avec celles des arbres syntaxiques (substitution et adjonction). Le pouvoir d'expression de cette construction n'est pas analysé, mais il me semble plus fort qu'un typage générique. Pour rester dans le domaine du football qui illustre le livre, le sujet de « gagner » est une équipe (c'est le typage générique de l'image de *gagne(m, eq)*) ; de plus, elle est gagnante d'un certain match qui est la variable principale de la DUDES de « gagner », ce qui introduit une relation avec ce match et augmentera la contrainte lors des unifications suivantes. Le texte ne précise toutefois pas comment l'inférence dans l'ontologie est utilisée dans les DUDES.

Reste, et ce n'est pas un mince problème, à rassembler les connaissances linguistiques utilisées pour l'interprétation, autrement dit les arbres élémentaires et leurs DUDES associés. Cette partie du travail est séparée en deux modules : la constitution d'un lexique syntactico-sémantique dont le format se veut indépendant du choix syntaxique, et la génération d'une grammaire dans le formalisme choisi. Le premier est assuré par le format de lexique Lemon : la version actuelle (2.0) repose sur une ontologie de 29 classes et 54 propriétés et fait aussi appel aux descriptions syntaxiques de l'ontologie Lexinfo, soit 192 classes, 156 propriétés et 271 individus permettant de noter une grande variété de propriétés syntaxiques[3]. On a là un format

2. DUDES signifie *Dependancy-based Underspecified DRS*.

3. Chiffres calculés à partir des ontologies publiques indiquées par le livre, juillet 2014.

de lexique bien plus détaillé que Skos. Le livre donne un ensemble d'exemples avec une réelle finesse linguistique : noms de classes et noms relationnels, verbes d'état, d'événement, résultatifs, adjectifs intersectifs, scalaires, modifieurs de classe, relationnels. Chacune des entrées lexicales fournies décrit des variantes morphologiques, un comportement syntaxique que je lis comme un cadre de sous-catégorisation, et un sens qui est un fragment de l'ontologie de domaine lié aux positions libres du cadre.

La génération de la grammaire à partir d'une de ces entrées produit plusieurs arbres selon les variantes morphologiques et les emplois : pluriel des noms, relations prépositionnelles, passif des verbes, comparatif et superlatif des adjectifs par exemple. D'après le schéma d'implémentation qui conclut le chapitre 5, l'automatisation repose sur des patrons adaptés à chaque cadre et aux arguments présents. J'ai compté quatre-vingt-quatre cadres dans Lexinfo – il n'était donc pas possible de décrire le dispositif en quelques pages.

Trois thèmes plus pointus concluent l'ouvrage : la désambiguïsation, le temps, le *Question Answering*. Une ontologie des intervalles temporels permet la modélisation du temps verbal (présent, passé, futur) et des expressions temporelles comme *aujourd'hui*, *hier*, *la semaine prochaine*. Je reviendrai sur la désambiguïsation dans l'analyse. La traduction de questions en requêtes SPARQL est une application sur des textes d'une phrase focalisés sur le contenu de la base de données de football.

Analyse

Le dispositif formel est cohérent et complet. Les données sont modularisées. De plus, elles sont toutes en RDF, connaissances du monde et données lexicales. Il est donc facile de les publier, de les analyser ou de les réutiliser. Il est aussi possible d'augmenter l'ontologie Lexinfo, d'ajouter un champ au lexique ou de modifier le moteur d'utilisation sans toucher au reste.

Les auteurs signalent eux-mêmes que l'expérimentation en grandeur nature est encore en projet. J'ai relevé quelques problèmes du passage à l'échelle. En premier lieu, l'inventaire des arbres syntaxiques élémentaires est contraint par la sémantique : « *The whole prepositional phrase corresponds to one atomic élément on the semantic side, so it should also be one atomic élément on the syntactic side* » (p. 40). Cela conduit à la multiplication des arbres syntaxiques élémentaires.

Le livre souligne aussi, à juste titre, la nécessité d'obtenir des représentations comparables pour des formulations différentes (p. 7-8) : « *an early opening goal by X* » et « *X opened the scoring shortly after kick off* » évoquent tous deux un match, un but, son auteur X et sa date. Si les fragments d'ontologie figurant dans les DUDES sont homologues pour des formulations différentes, c'est au prix d'une spécialisation poussée : les représentations sémantiques de *win* comme de *loose* évoquent un match. Cependant, le Webster indique que l'on peut aussi gagner un tournoi, une guerre, une récompense, sa vie, sans compter des sens plus éloignés. Ce thème est repris dans le chapitre sur la résolution des ambiguïtés : un élément peut avoir un choix de représentations sémantiques, celles-ci sont filtrées par la compatibilité des remplisseurs de leurs places syntaxiques trouvés dans le texte. On

gagne en pouvoir d'expression, mais assez peu en largeur de domaine, car chaque représentation est spécialisée.

Le choix de la variable principale des composants incluant un verbe semble aussi une difficulté : les exemples incluent des formes progressives, des passifs, des relatives, mais le degré de généralité de chaque exemple est difficile à apprécier. La marque du temps verbal s'unifie à la variable principale du verbe ; pour l'exemple de « gagner », il s'agit du match, mais je n'ai pas vu la solution pour « respecter », dont la représentation n'est pas ainsi réifiée et qui n'a pas de variable principale indiquée.

Conclusion

Le premier mérite de ce livre est d'allier résolument syntaxe et sémantique sans supposer une relation d'ordre entre elles. La modularisation très claire et la normalisation des données, le recours aux standards du Web sont très intéressants. Le tout repose sur l'idée, à mon sens pertinente, que la sémantique lexicale dépend du domaine et que la modularité permet l'adaptation. Il reste, et c'est je crois une condition inhérente au projet d'apprendre le lexique dont les auteurs font état, à calibrer le type de domaine pour lequel cette architecture passe à l'échelle.

Pierre-André BUVET. La Dimension lexicale de la détermination en français. *Honoré Champion*. 2013. 473 pages. ISBN 978-2-7453-2604-1.

Lu par **Catherine SCHNEDECKER**

LiLPa – Université de Strasbourg

L'ouvrage propose une typologie des déterminants du français fondée sur un inventaire de formes plus vaste que ce qui est traditionnellement préconisé, compte tenu d'une démarche globale appréhendant les phénomènes aux plans syntaxique, sémantique et même morphologique. Dans cette optique, le lexique occupe une place cruciale et il opère doublement, en tant que matériau constitutif de certaines sous-catégories de déterminants mais aussi en tant que tête lexicale du SN imposant divers niveaux de contraintes (syntaxiques et sémantiques) sur le déterminant. S'ensuit une nouvelle cartographie de la détermination qui oblige à repenser les frontières intercatégorielles.

L'ouvrage de Pierre-André Buvet résulte d'une HDR soutenue en 2009 à l'université de Paris 13. Pour autant, cet ouvrage ne se réduit pas aux aspects de la détermination, déjà nombreux, abordés par l'auteur. Il constitue au contraire d'abord une synthèse sur la détermination, comme le montre sa structure : en effet, l'ouvrage se subdivise en deux parties, plus une importante annexe qui fournit trois listes des principaux prédéterminants et antédéterminants du français, de l'ensemble des déterminants nominaux et de nombreuses séquences déterminatives figées.

La première partie de l'ouvrage comprend trois chapitres consistant en la définition et les propriétés de la détermination dans le cadre théorique auquel souscrit l'auteur. Le chapitre 2 inventorie les formes dites de la détermination simple *vs* complexe,

dont la caractérisation s'appuie sur d'abondantes batteries de tests. Le chapitre 3 aborde, comme l'indique son titre, l'ensemble des modifications du GN, dont une catégorie de modifieurs propositionnels (les complétives, infinitives et participiales), souvent occultés par les grammaires. Quant à la seconde partie, elle se consacre principalement aux déterminants défini et démonstratif, jusque dans leurs emplois anaphoriques. Est abordée également la modification des SN définis. Le chapitre 6 porte sur la détermination possessive, généralement peu abordée comme cela est rappelé à juste titre.

Il serait long et fastidieux d'entrer dans le détail de cette synthèse foisonnante par le nombre de déterminants étudiés ainsi que par l'abondance des tests, des exemples et des tableaux présentés tout au long d'une démarche extrêmement méthodique.

Nous présenterons simplement quelques-uns des points forts et originaux de ce travail, qui complètent une littérature déjà très fournie sur le sujet.

Un premier point fort de l'ouvrage a trait à l'ouverture théorique ainsi qu'à la multiplicité des travaux et des approches qui y sont exploités ou cités, allant de la sémantique formelle à la théorie des opérations énonciatives en passant par la grammaire générative.

Corollairement – et c'est un deuxième point fort – la détermination n'est pas seulement traitée sous un angle syntaxique : la sémantique, et même la morphologie, avec les questions de figement et de composition, sont également partie prenante de la démarche de l'auteur.

Ceci expliquant cela, la « portée » de la détermination, si l'on peut dire, s'en trouve considérablement modifiée. En effet, P.-A. Buvet démontre qu'elle ne concerne pas exclusivement la classe de ce qu'on a pu traditionnellement dénommer « articles », mais qu'elle englobe aussi des adverbiaux (*trop de*), des noms (*tonne, flopée, armée*) ou des séquences verbales (*je ne sais quel, n'importe quel*). Qu'à côté des déterminants standard existent des séquences déterminatives complexes, par exemple figées (*un méchant* dans *un méchant rhume*), et que les formes sémantiques de la détermination sont plus variées qu'il n'est dit généralement dans les grammaires, qu'elles soient intensives ou aspectuelles (*un début de, un_naissant*). De là aussi vient que la place des déterminants au sein d'un SN est autrement plus variable qu'il n'est généralement dit. Par ailleurs, ce que met en évidence cet ouvrage est l'interaction étroite entre la détermination et les propriétés des N actualisé : « *Les propriétés des noms déterminés sont le plus souvent des facteurs qui sont minimisés, voire totalement négligés dans les nombreux travaux qui leur sont consacrés* ».

Cela n'est pas nouveau dans la mesure où le facteur nominal ou lexical intervient, comme on sait, dans les oppositions traditionnelles entre massif et comptable ou concret et abstrait. Mais celles qui se manifestent ici engagent des sous-classes de N plus fines (les noms de vêtements, de sports, de maladies, d'affects et les noms humains, notamment) et, avec elles, un faisceau de contraintes extrêmement « sophistiquées ».

Bref, telle qu'elle ressort de cet ouvrage, la détermination apparaît de nature à renouveler profondément les descriptions lexicographiques, qui évoquent, par exemple, les emplois figurés de *tas* (*un tas d'ennuis*) ou *torrent* (*un torrent de larmes*) sans aller jusqu'à la détermination ou le contenu des grammaires, notamment scolaires. Dans le cadre méthodologique ambiant, elle pose aussi de sérieux problèmes pour les annotations dites de haut niveau ou l'identification automatique. Mais, plus largement encore, ainsi reconfigurées, les frontières de la détermination obligent à reconsidérer, peut-être même à redessiner, celles des catégories (adjectives, nominales, etc.) qui y sont impliquées. C'est dire si le sujet intéresse non seulement les linguistes de tous bords, mais aussi les didacticiens, les informaticiens ou encore les traducteurs.

Laurent GOSSELIN, Yann MATHET, Patrice ENJALBERT, Gérard BECHER. Aspects de l'itération. L'expression de la répétition en français : analyse linguistique et formalisation. *Peter Lang*. 2013. 372 pages. ISBN 978-3-0343-1415-2.

Lu par **Natalia Grabar**

UMR 8163 Savoirs, Textes, Langage STL – Université de Lille *3*

L'ouvrage regroupe trois contributions dédiées à la formalisation de l'itération temporelle. Il s'agit d'un ouvrage interdisciplinaire, où les auteurs, les chercheurs en linguistique (sémantique formelle), informatique et logique, se proposent de modéliser la notion de l'itération temporelle. Comme la réflexion a été faite dans le cadre d'un projet, les chercheurs se focalisent sur des notions identiques ou proches et il existe également des renvois entre les contributions. Le travail est effectué avec des exemples langagiers, réels ou jouets. En revanche, le lien avec le TAL et le traitement automatique de corpus est absent excepté la mention d'une thèse d'informatique non présentée dans l'ouvrage.

L'itération temporelle est la répétition dans le temps d'un même procès, en sachant que la *répétition dans le temps* sous-entend que les intervalles de procès correspondants ne coïncident pas (il y a au moins une succession des bornes initiales de ces intervalles) et qu'un *même procès* sous-entend, quant à lui, qu'un événement unique corresponde à chaque occurrence de ce procès. La modélisation de cet objet a été donc soumis à trois disciplines. En plus de la modélisation, chaque contribution propose également une visualisation des procès et de leurs itérations. Les modélisations proposées par chaque discipline impliquée se fondent sur les travaux existants tout en proposant des développements nouveaux.

Dans la première contribution dédiée au modèle linguistique, il est indiqué que l'on distingue traditionnellement deux aspects de l'itération : l'aspect lexical (marqué par les lexèmes verbaux, qui permettent de construire le procès) et l'aspect grammatical (exprimé par les conjugaisons de ces verbes, qui permettent d'exprimer la façon de voir ce procès du point de vue aspectuel). Ici, l'auteur propose d'abandonner cette dichotomie du fait qu'elle ne représente pas correctement la situation en français et n'est pas généralisable à d'autres langues. L'auteur propose

plutôt une opposition sémantique en distinguant l'aspect conceptuel (fondé sur le processus sémantico-cognitif de construction des procès par catégorisation) et la visée aspectuelle (opérant sur la présentation des procès préalablement construits par l'intermédiaire des intervalles). Ceci permet d'avoir un point de vue plus global et complet des procès et de leurs itérations. Parmi d'autres notions linguistiques analysées se trouvent, par exemple, les types de procès (état, activité, accomplissement et achèvement), les phases de procès (préparatoire, initiale, médiane, finale et résultante), les visées aspectuelles de procès (aoristique, inaccomplie, accomplie et prospective), la coupure modale (faisant la distinction entre les procès réalisés et irrévocables et les procès futurs et possibles), les types de compléments circonstanciels temporels (de durée, comme *pendant deux heures, en une semaine*, et de localisation temporelle, comme *en 2014, un jour, lorsque je lisais l'ouvrage*), la portée des circonstanciels, etc. Un autre point important de ce modèle est que l'itération des procès est envisagée de manière compositionnelle, où le calcul de la sémantique globale d'un procès résulte de la sémantique de ses composants. Cependant, l'auteur propose que la compositionnalité soit non pas atomique, mais holiste, où la sémantique des procès est vue dans leur globalité et respecte autant la sémantique des marqueurs individuels que les relations entre ces marqueurs et les connaissances encyclopédiques et pragmatiques. De la même manière, ce modèle permet d'agglomérer les procès et de résoudre les conflits, comme dans le cas de trois marqueurs impliqués dans la phrase « *Depuis deux mois, il mangeait en 10 minutes* » *:* imparfait *mangeait* inaccompli, *en 10 minutes* aoriste et *depuis deux mois* compatible uniquement avec les aspects accomplis et inaccomplis. Le modèle qui s'inscrit dans les travaux en informatique est fondé sur les principes de la programmation objet et des espaces mentaux. Ainsi, les itérations de procès sont caractérisées, par exemple, par les objets (qui correspondent aux procès) et les classes d'objets, les relations entre les classes (associative, d'héritage), les itérateurs déclenchés par les marqueurs linguistiques, les relations générales entre les procès (temporelle, causale et méronimique), les relations temporelles entre les procès (concomitance et succession), la récursivité comme le mécanisme de l'itération, et la sélection pour introduire des contraintes (conditions, exceptions).

Le troisième modèle de l'itération repose sur les principes algébriques et logiques, avec une attention particulière portée à la quantification et la pluralité nominale. Les postulats logiques sont posés et permettent de prendre en charge les notions comme le temps, l'intervalle, les restrictions, les ensembles, les composants ou les relations (hiérarchie, inclusion, mesure et topologie). Comme le travail est effectué avec des données langagières, une description spécifique est proposée pour les différents types de quantifieurs : les déterminants (*e.g., les, tous les, chaque, un, un certain, la plupart des, presque tous les, certains, quelques*), les expressions de quantification explicite (*e.g., trois jours par mois, trois fois par jour, deux mois sur douze*), et les compléments (les heures, les intervalles).

Les chercheurs s'attaquent à des notions liées à l'itération de procès assez complexes à décrire et à représenter. Les notions abordées peuvent être proches (agglomérat, modèle, ensemble et série de procès) ou identiques (quantification, relations entre les éléments itérés, etc.).

Cet ouvrage peut être intéressant pour les chercheurs en TAL travaillant sur la temporalité. Bien que le lien avec les travaux autour de la détection des expressions temporelles, de la norme TimeML et de la construction de lignes de vie et de lignes temporelles ne soit pas établi, les modèles proposés dans l'ouvrage peuvent aider dans le calcul et la représentation des lignes temporelles et des relations entre les procès. De même, les auteurs donnent quelques pistes pour la résolution de conflits entre les représentations sémantiques des expressions temporelles.

Kevin BRETONNEL COHEN, Dina DEMNER-FUSHMAN. Biomedical Natural Language Processing. *John Benjamins publishing company*. 2013. 160 pages. ISBN 978-9-0272-4998-2.

Lu par **Thierry HAMON**

Université Paris-Nord - LIMSI – UPR 3251 – Orsay

Cet ouvrage présente l'état de l'art du traitement automatique des textes biomédicaux en anglais, c'est-à-dire de la littérature scientifique produite par les biologistes et les médecins, ainsi que des textes cliniques décrivant le parcours de santé des patients hospitalisés. Ce livre est destiné aux chercheurs en TAL qui désirent s'intéresser à l'analyse de ce type de documents textuels et passe en revue les différents contextes applicatifs dans lesquels le TAL peut être mis en œuvre. L'ouvrage est composé de onze chapitres assez courts dédiés à la résolution d'un problème applicatif particulier (extraction d'information, recherche d'information, etc.). À l'exception des deux premiers, chaque chapitre est structuré de manière similaire : la thématique abordée est justifiée du point de vue biomédical ; les problématiques et les difficultés liées à la thématique, au domaine et aux types de textes sont mises en avant ; certains travaux ou outils dédiés à cette thématique sont décrits avec plus ou moins de détails.

Le premier chapitre est une courte introduction rappelant les notions de TAL pertinentes pour le traitement de corpus biomédicaux. C'est aussi l'occasion pour les auteurs d'introduire les types de données textuelles disponibles dans le domaine et d'évoquer les difficultés liées à l'analyse de ces textes et les solutions généralement mises en œuvre.

Dans le chapitre 2, les auteurs font un historique du TAL dans le domaine biomédical. Ils présentent ainsi les outils marquant les évolutions majeures du domaine, mais aussi les ressources, les corpus et les portails disponibles (ces portails étant souvent des sources pour la constitution de corpus textuels). Le chapitre se termine par une description des problèmes éthiques et légaux liés aux données cliniques.

Le chapitre 3 est consacré à la reconnaissance d'entités nommées dans les textes biomédicaux. Après avoir montré l'importance de cette tâche dans le domaine, mais aussi les difficultés à identifier les entités nommées dans les textes (à la fois pour des raisons d'ambiguïté, de polysémie et de métaphore), les auteurs présentent deux systèmes typiques : l'un s'appuyant sur des règles pour reconnaître des noms de gènes, l'autre utilisant des informations statistiques pour apprendre des patrons

d'identification des noms de maladies. Le chapitre se conclut sur la manière d'évaluer les approches proposées et les collections de données disponibles pour une telle évaluation.

Le chapitre 4 s'intéresse à l'extraction de relations sémantiques. Les auteurs se placent d'emblée dans une perspective d'extraction d'information. L'extraction des relations est considérée comme le moyen de remplir des formulaires permettant de décrire une maladie ou un gène avec des informations associées. Ainsi, pour identifier des réseaux génomiques à partir de textes de biologie ou pour décrire cliniquement les patients à travers la fouille de textes cliniques, les méthodes présentées s'appuient aussi bien sur des méthodes par apprentissage que sur des approches à base de règles pour répondre à cette problématique. Comme le chapitre précédent, ce chapitre se termine par une évocation de problématiques liées à l'évaluation.

Le chapitre 5 est consacré à la recherche d'information, principalement dans la littérature biomédicale. Une première partie du chapitre décrit la base PubMed/Medline sur laquelle portent de nombreux travaux. La recherche d'information dans les textes biomédicaux s'appuyant naturellement sur les connaissances du domaine pour réduire les problèmes de polysémie et d'ambiguïté sémantique, le métathésaurus UMLS, qui regroupe plus d'une centaine de ressources terminologiques biomédicales, est considéré comme une ressource primordiale dans le domaine biomédical. Cependant, l'utilisation de l'UMLS n'est pas une solution à ces problèmes : pris dans son ensemble, l'UMLS contient lui-même des termes polysémiques ou ambigus étant donné qu'il a été constitué à partir de nombreuses ressources terminologiques. Des travaux se sont donc intéressés à l'exploitation des informations issues de l'UMLS pour améliorer la recherche d'information (identification des concepts de l'UMLS pertinents pour les requêtes, utilisation de synonymes). Les auteurs abordent aussi cette problématique dans le contexte plus particulier de l'interrogation de bases de connaissances biologiques et de la désambiguïsation des noms de gènes. Enfin, les auteurs présentent quelques travaux liés à la mise en œuvre de systèmes de recherche sur du texte plein, des images, des figures et des légendes de figures.

Dans la continuité du chapitre précédent, le chapitre 6 aborde la problématique de la normalisation de concepts, utile à la fois en extraction et en recherche d'information. Deux aspects sont présentés : la normalisation de noms de gènes et la normalisation de termes issus notamment de textes cliniques. Dans le premier cas, il s'agit d'associer les variantes de noms de gènes que l'on peut trouver dans la littérature scientifique biomédicale, avec la forme normalisée présente dans la base « Entrez Gene » (base recensant les informations liées aux gènes). Après avoir présenté la problématique, les auteurs décrivent les solutions possibles et notamment la mise en œuvre du système GNAT. Dans le second cas, il s'agit de la normalisation de termes issus des textes cliniques, et de l'identification des concepts de l'UMLS. Cela nécessite la reconnaissance de termes associés aux concepts mais aussi de leurs variantes malgré les difficultés liées aux ambiguïtés et à la polysémie. Le système MetaMap développé par la NLM (National Library of Medicine) y est décrit. Les auteurs présentent également ses limites et ses évolutions possibles.

Le chapitre 7 s'inscrit dans la continuité du précédent. Après avoir décrit en détail l'UMLS et Gene Ontology, les auteurs montrent comment le TAL peut être utilisé pour reconnaître dans les textes, les termes issus de ces deux ressources, pour vérifier la qualité des ressources (pour ajouter des relations manquantes par exemple), pour aligner des ontologies ou pour les mettre en relation. Ici aussi, les auteurs illustrent leur propos en décrivant en détail des méthodes dédiées.

Le chapitre 8 s'intéresse au résumé automatique de textes biomédicaux à travers la description d'un système de résumé multidocument destiné à synthétiser la littérature médicale. L'état de l'art de cette thématique se poursuit par la présentation de systèmes dans le domaine de la génomique pour enrichir des bases de données avec des connaissances sur les gènes, c'est-à-dire des descriptions en langue naturelle, des fonctions des gènes ou des interactions de protéines.

Le chapitre 9 est consacré aux systèmes de question-réponse. Après un rappel des principes de base de ces systèmes et des problématiques particulières concernant l'interrogation de la littérature médicale (notamment dans le cadre de la médecine fondée sur les faits – *evidence based medicine*) et celle d'articles de génomique, les auteurs décrivent en détail les différentes étapes de mise en œuvre d'un système dédié à l'interrogation de la base PubMed.

Le chapitre 10 est une réflexion originale sur les méthodes de génie logiciel devant être mises en place dans le contexte du TAL et sur les problématiques d'ingénierie, en particulier pour l'analyse de la qualité des outils de TAL dans le domaine biomédical. La première partie du chapitre est un rappel des techniques classiques de vérification de code, tandis que la seconde partie est consacrée à la mise en œuvre de ces techniques lors du développement d'approches ou d'applications de TAL.

Dans le dernier chapitre, les auteurs décrivent succinctement la problématique de la constitution et l'annotation de corpus dans le domaine biomédical. La majorité du chapitre est consacré à la description des corpus annotés disponibles dans le domaine. On peut regretter qu'une partie du chapitre ne soit pas consacrée à la définition de guide de définition ou d'annotation de corpus.

Liste des auteurs